ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Model for Prediction of Dropout Student Using ID3 Decision Tree Algorithm

[I]Sweta Rai, [II]Priyanka Saini, [III]Ajit Kumar Jain
[I,II]M.Tech in Computer Science, Dept. of AIM & ACT, Banasthali University, Jaipur, Rajasthan, India
[III]Assistant Professor, Dept. of AIM & ACT, Banasthali University, Jaipur, Rajasthan, India
E-mail : swetarai90@gmail.com

Abstract
The objectives of this research work is to identify relevant attribute from socio-demographic, academic and institutional data of first year students from undergraduate at the University and design a prototype machine learning tool which can automatically recognize whether the student continue their study or drop their study using classification technique based on decision tree. For powerful decision making tool different parameter are need to be considered such as socio-demographic data, parental attitude and institutional factors. The generated knowledge will be quite useful for tutor and management of university to develop policies and strategies related to increase the enrolment rate in University and to take precautionary and advisory measures and thereby reduce student dropout. It can also use to find the reasons and relevant factors that affect the dropout students.

Keywords
Predicting dropout student, machine learning algorithm, classification, decision tree

I. Introduction

We live in the information- era, accumulating data is easy and storing it inexpensive. Today the amount of stored information increases day by day in all areas. Unfortunately, as the amount of machine readable information increases, the ability to understand and make use of it does not keep pace with its growth. Mining novel and most useful information from huge amount of data is known as data mining. Data mining techniques can be applied in different application domain, such as, Education, Banking, Marketing, Fraud detection and Telecommunications. It is an inter- disciplinary field involving concepts from Artificial intelligence, Machine Learning, Neural networks, Database technology, Statistics, Mathematics, Clustering and Data visualization. Machine learning provides tools for analysis of large quantities of data automatically such as feature selection. Feature selection is to choose a subset of input data most useful for analysis and future prediction by eliminating features, which are irrelevant or of no predictive information. Feature selection is use for increasing the predictive accuracy and reducing complexity of learner results [1].

The earlier prediction of dropout student is challenging task in the higher education. Data analysis is one way to scale down the rate of dropout students and increase the enrollment rate of students in the university. It is fact that student dropout quite often in the first year of graduation. Dropout in residential university is caused by academic, family and personal reasons, campus environment and infrastructure of university and varies depending on the educational system adopted by the university. Thus, this study is quite useful for better planning and implementation of education program and infrastructure to increase the enrollment rate of students in particular courses provided by the university.

The main aim of this paper is to design a classification model using decision tree induction algorithm and classifier rules to predict whether student will graduate or not using the historic data. In this paper, ID3 decision tree algorithm is used to design a conceptual model. Information like age, parent's qualification, parent's occupation, academic record, attitude towards university was accumulated from the student's residing in university campus, to predict list of students who need special attention.

II. Literature Review

Educational data mining is an emerging application of data mining. Many researcher and authors have explored and discussed various application of data mining in higher education.

Romero and Ventura [3] perform comprehensive study of Educational data mining from 1995 to 2005.

Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, and M. Inayat Khan [4] applied K-mean clustering to analyze learning behavior of students which will help the tutor to improve the performance of students and reduce the dropout ratio to a significant level.

D'Mello [5] studied on bored and frustrated student. Romero [6] studied on the factors that predict failure and non-retention in college courses.

III. Pattern Discovery Process For Student Dropout

Data mining is also referred as knowledge discovery in databases (KDD), this is a technique for discovering a hidden pattern, previously unknown and potentially useful information from data. This process involves various steps such as preprocessing, make data mining and visualize the outcome [2].

In the process of discovering patterns, the following steps were performed:

A. Selection Step

The aim of this step is to select data set from internal or external sources of data. Marks obtained by students in secondary and higher secondary school were selected from Academic control office of the university and personal record or their view for university were collected through survey based on questionnaire. From 300 records, only 220 data of student belonging to first year with most relevant attribute were chosen to this study. The outcome of 220 records belonging to socio-demographic, academic, and institutional factors. These data were stored in file "studentdropout. csv" using MS-Excel.

B. Pre-processing Step

The main aim of this stage is to obtain preprocessed data in order to retrieve high quality patterns. The preliminary survey of student at University includes data about 300 students, described by 41 parameters including ID, Age, Date of birth, category, marital

status, residence, state of domicile, mother tongue, religion, family type, family annual income, father's education, mother's education, father's occupation, mother's occupation, location of school, percentage in secondary school, percentage in higher secondary, medium of education, stream in higher secondary, course admitted, admission type, course satisfaction, syllabus of course, whether parents are interested in their children higher education, reason for selecting the university, consent of parents to work private sector, family experiences stress, source of information to know about the university, join the university by your choice, like university or not, educational system of the university, infrastructure of university, extra-curriculum activities in university, entertainment in university, attendance in class, time for self study, participation in extra-activity, cope with pressure at university, placement in university and dropout status shows whether the student will withdraw from course or not, etc. the provided data need many transformations. Some of parameters are removed such as ID, age, date of birth, category, marital status, state of domicile, mother tongue, religion, the gender field containing only one value- female, because university concerns only female students, the marital status field containing one value- unmarried, etc. A categorical variable is constructed based on the numeric parameter percentage in secondary and higher secondary school. A grade scale is used for evaluation of student performance at school. "Garde A" students are considered those who have a percentage greater than 85, "Grade B"- in the range between 75 and 85, "Grade C"- in the range between 65 and 75, and "Grade D" in the range below 65. A four level scale is used in the family annual income. "VHigh" annual income are considered those who have income greater than 6 lakhs, "High" annual income range between 4 lakhs and 6 lakhs, "Medium" annual income range between 2 lakhs and 4 lakhs and "Low" annual income in the range below 2 lakhs. A categorical target variable "Dropout status" is constructed based on the view of respondents; it has two possible values- "Yes" (students who are completely decided to withdraw from their course) and "No" (students who are want to continue their study).

The final dataset used for the study contains 220 instances (183 in the "No" category and 37 are in "Yes" category) each described with 34 attributes (1 output and 33 input variables), nominal and numeric. The attribute related to socio-demographic data include age, category, residence, family type, parents qualification and parents occupation etc. Location of school, medium of education, academic performance, and stream in higher secondary school attributes referring to the student's pre characteristics. There are some attributes describing university features such as education system, infrastructure, and extra-curriculum activities. The study is limited to the student data for undergraduate. The sample contains data about female student.

At this stage, in order to generate knowledge the 12 most representative attribute were selected from database based on correlation feature selection method is shown in Table 1.

C. Data Mining Step

The aim of this step is to extract a hidden pattern from huge amount of data by applying intelligent task such as classification, clustering and association etc. For discovering student dropout pattern, classification using decision tree technique was used. Classification is a supervised learning. It is two-step process. In first step, model is built using historical data. In second step, the model is used to classify future test data for which the class levels are unknown [3].

Decision tree one of the most popular model, because it is simple and easy to understand. It is flow- chart like tree structure, where each internal node denotes test on an attribute, each branch represents an outcome of the test, and leaf node represent class. For building a model that would classify the students into the two classes, depending on the historical data. WEKA data mining tool are used in the experimental study, collection of machine learning algorithms for data mining tasks and multiple tools for data pre-processing, classification, regression, clustering, association rules and visualization. In this paper, Weka software is used for feature selection including a decision tree algorithm ID3.

To build a decision tree ID3 employ a greedy approach (information theoretic measure) and correlation feature selection (CFS).

D. Evaluation Step

Evaluation is the final step of KDD process, interpret the result.

IV. Feature Selection

In this paper, use feature selection for supervised machine learning tasks on the basis of correlation between features. It contains a good feature subset that is highly correlated with class otherwise it is irrelevant. Correlation measures the strength of linear association between two variables. The range of correlation is -1.0 and +1.0. If the correlation is positive, the relation is positive. If it is negative, the relation is negative. Feature selection has two approaches forward selection and backward selection. It selects a subset of input variables by eliminating irrelevant features. In this paper, Correlation-Based Feature Selection (CFS) is use to find the feature subsets that are highly correlated with the class but minimal correlation between features combined with search strategy best-first search (BFS). CFS measures correlations between nominal features, so numeric features are first discretized. BFS starts with empty set of features and generate all possible single feature expansions. The subset with highest evaluation is chosen and expanded in the same manner by adding single features. If expanding a subset results in no improvement, the search back to the next best unexpanded subset and continues from there.

Equation for CFS:

$$M_s = \frac{K\overline{r_{cf}}}{\sqrt{K + K(K-1)\overline{r_{ff}}}}$$

Where M_s is the heuristic "merit" of feature subset S containing K features, r_{cf} is the mean feature-class correlation and r_{ff} is the average feature-feature inter-correlation.

The search begins with the empty set of features, which has zero merit. The subset with highest merit found during search the search is used to reduce the dimensionality of training and test data. The search will terminate if five consecutive fully expanded subsets show no improvement over the current best subset. Reduced datasets may be passed to machine learning (ML) for building a classifier to predict the dropout student see in Figure 1.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

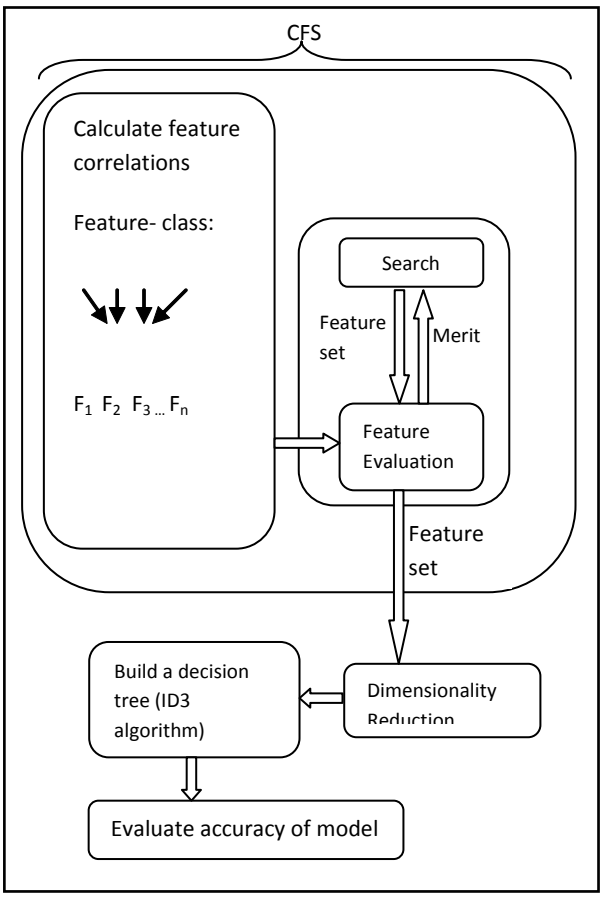


Fig. 1: Methodology of proposed work

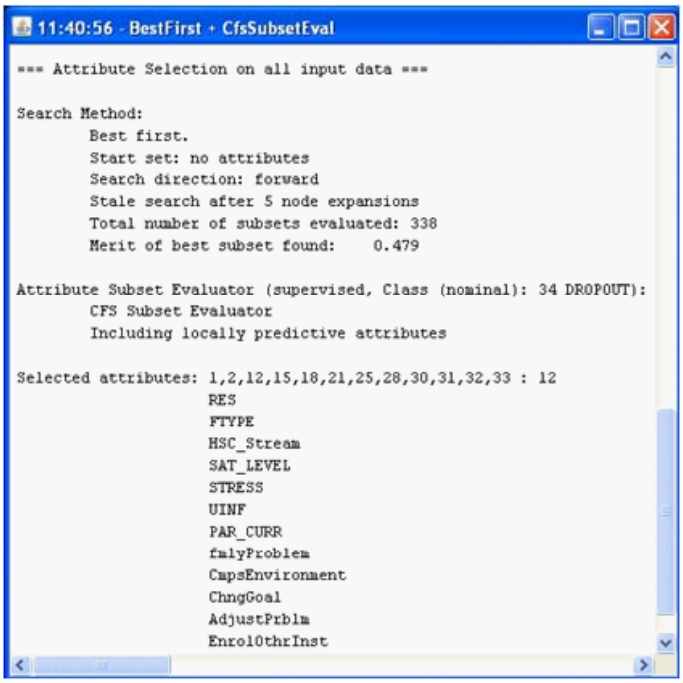


Fig.: Attribute selected using Correlation feature selection

V. Background Of ID3

ID3 (Iterative Dichotomizer 3) algorithm is invented by J. Ross Quinlan in 1979. It is used for building the decision tree using information theory invented by in 1948. It builds the decision tree from top down, with no backtracking. Information Gain is used to select the best attribute for classification.

A. Entropy

It is a measure of uncertainty about a source of message. It ranges from 0 to 1. When entropy is 1 means dataset is homogenous. Entropy is calculated by formula:

$$Entropy(s) = \sum_{i=1}^{c} -P_i Log_2 P_i$$

Where, P_i is the probability of S belonging to class i.

B. Information Gain

It measures the expected reduction in entropy. ID3 calculates the Gain of all attributes, and select the one with highest gain. To calculate Gain use formula:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where Values(A) is the set of all possible values for attribute A, and S_v is the subset of S for which the attribute A has value v. The attribute with highest information gain among the attributes are located as root node in the decision tree.

C. Implementation of ID3 Algorithm

Step 1: compute classification entropy.
Step 2: for each attribute, calculate information gain using classification attribute.
Step 3: select attribute with highest information gain.
Step 4: remove node attribute, for future calculation.
Step 5: repeat steps 2-4 until all attribute have been used.

```
Function ID3 (Input attribute, Output attribute, Training data)
{
If (Training data is empty)
   {
  Return a single node with Failure;
   }
If (all records in training data have positive value)
   {
Return a single node with level positive.
   }
If (all records in training data have negative value)
   {
 the single node with level negative;
   }
If (input attribute is empty)
   {
Return a single node with the value of the most frequent value;
   }
Otherwise
{
Compute information gain for each attribute;
Split the attribute with highest information gain value;
Return a tree with root node X and arcs X1, X2…, Xm;
Recursively call the ID3 function until all attribute have been used.
}
}
```

VI. Empirical Study

The aim of this study is to select a highly correlated feature which are associated with dropout student and design a classification model for future prediction of whether student will drop the course or continue and study the cause of dropout students.

VII. Experiment Setup

This study uses a data set of 220 student collected randomly through survey at University. The data set contains 37 positive and 183 negative reviews collected from questionnaire with 34 variables. In this experiment, CFS based on best first search was used with five nodes expansion to reduce the dimension of data set by removing the uncorrelated attribute to the prediction class. In addition to feature selection algorithm, also select ID3 decision tree algorithm, to evaluate the accuracy on selected features.
The experiments are conducted using Weka tool. Run a CFS with search space best-first search on dataset and record the selected features. Then apply ID3 only on the selected features set and record overall accuracy by 10-fold cross- validation. To evaluate the performance of classification, this paper has adopted precision, recall and F-measure as a performance measure.

$$True\ positive = \frac{TP}{P}$$

$$False\ Negative = \frac{FP}{N}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

VIII. Result and Discussion

CFS with search strategy best-first implemented on the input data, it could be seen that subsets of 12 attributes was selected out of 33 attributes (Table 1). Then the ID3 decision tree algorithm is employed on selected subset of features and record using 10 cross-validation (Fig. 1). Attribute with highest information gain is used as a root node (Table 2). To study the actual and predicted class a confusion matrix table was used (Table 3). The dropout data set is classified into two groups Yes and No based on this 2x2 confusion matrix for id3 was constructed shows accuracy percent 98.1818 for ID3. It indicates that it is the best classifier for predicting the student who will dropout or not at the University.

Table 1: Feature Selection based on Correlation Feature Selection

Variables	Description	Selected by (CFS) Correlation feature selection
Res	Residence	Yes
FType	Family Type	Yes
FAIn	Family annual income	
FEdu	Father's Education	
MEdu	Mother's Education	
FOcc	Father's Occupation	
MOcc	Mother's Occupation	
SLoc	School location of student	
HSG	Students grade / percentage in High School (10th)	
SSG	Students grade / percentage in Senior Secondary (12th)	
Med	Medium of education	
HSC_Stream	Stream in Senior Secondary(12th)	Yes
CAdm	Course Admitted	
AType	Admission Type	
SAT_Level	Satisfaction with course	Yes
CSyllabus	Syllabus of Course	
UExpenses	Parents meet the University Expenses	
Stress	Family experiences stress	Yes
LikeUni	Like this university	
EduU	Educational system of the university	
UINF	Infrastructure of university	Yes
ActU	Extra-Curriculum activities in university	
EntertU	Entertainment in university	
SelfStudy	Time for self study	
Par_Curr	Participate in extra curriculum activity	Yes
Placement Status	Placement status	
TSRelation	Teacher student relationship	
fmlyProblem	Family problem	Yes
Hsickness	Home sickness	
cmpsEnv	Campus Environment	Yes
chngGoal	Change of goal	Yes
adjustPrblm	Adjustment Problem	Yes
EnrolOthrInst	Enrolled in other institute	Yes

A. Step by Step Calculations

Step 1: the dataset D of 220 instances with 37 "Yes" and 183 "No".

$$Entropy(D) = -\left(\frac{37}{220}\right)Log_2\left(\frac{37}{220}\right) - \left(\frac{183}{220}\right)Log_2\left(\frac{183}{220}\right) = 0.653529$$

Step 2: Attribute Residence

Residence value can be Rural and Urban
Residence= rural is of occurrences 51
Residence= urban is of occurrences 169
Residence= rural, 2 of instances are 'yes' and 49 are 'no'

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Residence= urban, 35 of instances are 'yes' and 134 are 'no'
Entropy (D rural) = -(2/51)log2(2/51)-(49/51)log2(49/51)
= 0.238685
Entropy (D urban) =
- (35/169)log2(35/169)-(134/169)log2(134/169) = 0.735904
Gain (D, residence)= Entropy(D)-(51/220)*Entropy(D rural) – (169/220)*Entropy(D urban)
=0.653529-(51/220)*0.238685-(169/220)*0.735904
=0.032889

Step 3: Attribute stress

Stress value can be No stress, financial, illness, other
Stress= No stress is of occurrences 129 (129 instances are 'no')
Stress= financial is of occurrences 48 (8 instances are 'yes' and 40 are 'no')
Stress= illness is of occurrences 16 (9 instances are 'no' and 7 are 'yes')
Stress= other is of occurrences 27 (22 instances are 'yes' and 5 are 'no')
Entropy (D no)=-(0/129)log2(0/129)-(129/129)log2(129/129) =0
Entropy (D financial) = -(8/48)log2(8/48)-(40/48)log2(40/48)
= 0.650022
Entropy (D illness) = -(7/16)log2(7/16)-(9/16)log2(9/16)
= 0.988699
Entropy (D other) = -(22/27)log2(22/27)-(5/27)log2(5/27)
= 0.69129
Gain (D, stress) =
0.653529-(48/220)*0.650022-(16/220)*0.988699-(27/220)*0.69129
= 0.354961

Step 4: Attribute family type

Ftype value can nuclear and joint
Ftype= nuclear is of occurrences 115 (5 instances are 'yes' and 110 are 'no')
Ftype= joint is of occurrences 105 (32 instances are 'yes' and 73 are 'no')
Entropy (D nuclear) =
-(5/115)log2(5/115)-(110/115)log2(110/115)= 0.258019
Entropy (D joint) =
-(32/105)log2(32/105)-(73/105)log2(73/105) = 0.887034
Gain (D, Ftype) =
0.653529-(115/220)0.258019-(105/220)0.887034 =0.095298

Step 5: Attribute HSC_Stream

HSC_Stream value can be math, arts(math), bio, com and arts
HSC_Stream= math is of occurrences 179 (21 instances are 'yes' and 158 are 'no')
HSC_Stream= arts(math) is of occurrences 8 (8 instances are 'no')
HSC_Stream= bio is of occurrences 5 (4 instances are 'yes' and 1 is 'no')
HSC_Stream= commerce is of occurrences 20 (4 instances are 'yes' and 16 are 'no')
HSC_Stream= arts is of occurrences 8 (8 instances are 'yes')
Entropy (D math) = 0.521603
Entropy (D arts(math)) = 0
Entropy (D bio) = 0.721928
Entropy (D com) = 0.721928
Entropy (D arts) = 0
Gain (D, HSC_Stream) = 0.147097

Step 6: Attribute Satisfaction Level

Sat_Level value can be v.satisfied, satisfied and not satisfied
Sat_Level= V.satisfied is of occurrences 57 (57 instances are 'no')
Sat_Level= satisfied is of occurrences 122 (17 instances are 'yes' and 105 are 'no')
Sat_Level= not satisfied is of occurrences 41 (20 instances are 'yes' and 21 is 'no')
Entropy (D v.satisfied) = 0
Entropy (D satisfied) = 0.582519
Entropy (D not satisfied) = 0.999571
Gain (D, Sat_Level) = 0.144212

Step 7: Attribute Enrolled on other institute (EnrollOthrInst)

EnrollOthrInst value can be yes and no
EnrollOthrInst = yes is of occurrences 19 (14 instances are 'yes' and 5 are 'no')
EnrollOthrInst =no is of occurrences 201 (23 instances are 'yes' and 178 are 'no')
Entropy (D yes) = 0.831477
Entropy (D no) = 0.513129
Gain (D, EnrollOthrInst) = 0.112907

Step 8: Attribute change of goal (ChngGoal)

ChngGoal value can be yes and no
ChngGoal = yes is of occurrences 15 (12 instances are 'yes' and 3 are 'no')
ChngGoal =no is of occurrences 205 (25 instances are 'yes' and 180 are 'no')
Entropy (D yes) = 0.721928
Entropy (D no) = 0.534944
Gain (D, ChngGoal) = 0.105836

Step 9: Attribute like campus environment (CmpsEnvironment)

CmpsEnvironment value can be yes and no
CmpsEnvironment = yes is of occurrences 33 (18 instances are 'yes' and 15 are 'no')
CmpsEnvironment =no is of occurrences 187 (19 instances are 'yes' and 168 are 'no')
Entropy (D yes) = 0.99403
Entropy (D no) = 0.474061
Gain (D, CmpsEnvironment) = 0.101473

Step 10: Attribute like participation in extra-activity (Par_Curr)

Par_Curr value can be yes and no
Par_Curr = yes is of occurrences 144 (4 instances are 'yes' and 140 are 'no')
Par_Curr =no is of occurrences 76 (33 instances are 'yes' and 43 are 'no')
Entropy (D yes) = 0.183122
Entropy (D no) = 0.987475
Gain (D, Par_Curr) = 0.19254

Step 11: Attribute adjustment problem in hostel (AdjustPrblm)

AdjustPrblm value can be yes and no

AdjustPrblm = yes is of occurrences 15 (9 instances are 'yes' and 6 are 'no')
AdjustPrblm =no is of occurrences 205 (28 instances are 'yes' and 177 are 'no')
Entropy (D yes) = 0.970951
Entropy (D no) = 0.575226
Gain (D, AdjustPrblm) = 0.051322

Step 12: Attribute family problem (fmlyprblm)

fmlyPrblm value can be yes and no
fmlyPrblm = yes is of occurrences 60 (19 instances are 'yes' and 41 are 'no')
fmlyPrblm =no is of occurrences 160 (18 instances are 'yes' and 142 are 'no')
Entropy (D yes) = 0.90072
Entropy (D no) = 0.507411
Gain (D, fmlyPrblm) = 0.038852

Step 13: Attribute university infrastructure (UINF)

UINF value can be yes and no
UINF = Excellent is of occurrences 30 (3 instances are 'yes' and 27 are 'no')
UINF = V.Good is of occurrences 59 (7 instances are 'yes' and 52 are 'no')
UINF = Good is of occurrences 113 (15 instances are 'yes' and 98 are 'no')
UINF = Poor is of occurrences 18 (12 instances are 'yes' and 6 are 'no')
Entropy (D excellent) = 0.468996
Entropy (D v.good) = 0.525451
Entropy (D good) = 0.564914
Entropy (D poor) = 0.918296
Gain (D, UINF) = 0.083365

Step 14: Find which attribute is the root node and rank the attribute with corresponding information gain shown in Table 2.

Gain (D, Stress) = 0.354961 is highest.
Therefore, "Stress" attribute is root node in the decision tree. "Stress" as root node has four possible values- no stress, financial, illness and other.

Table 2: Ranked Attribute with respect to Information Gain

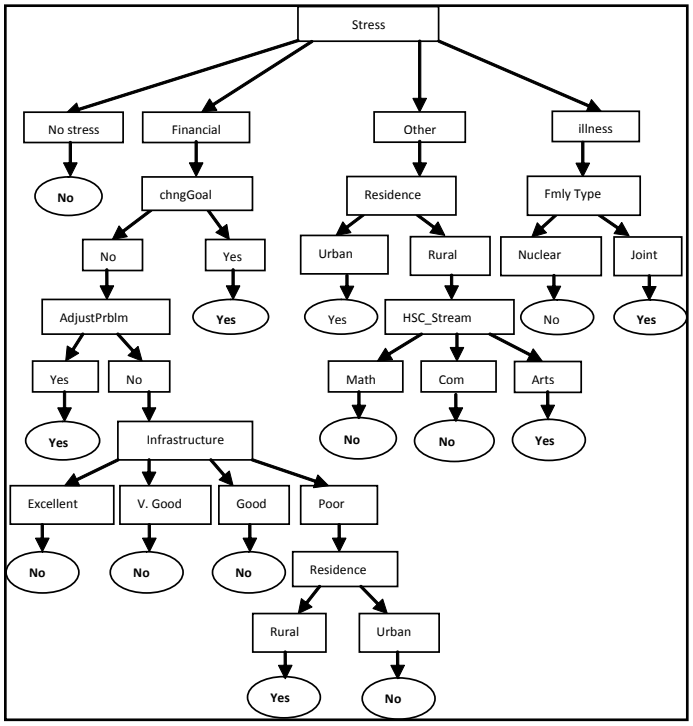
Information Gain	Attribute
0.355	STRESS
0.1925	PAR_CURR
0.1471	HSC_Stream
0.1442	SAT_LEVEL
0.1129	EnrolOthrInst
0.1058	ChngGoal
0.1015	CmpsEnvironment
0.0953	FTYPE
0.0834	UINF
0.0513	AdjustPrblm
0.0389	fmlyProblm
0.0329	RES

Step 15: Find which attribute is the next decision node.

"Stress" has four possible values.
So root node has four branches (no, financial, illness, other).
Stress = no stress is of occurrences 129 (129 instances are no)
Entropy (Stress No) = -(129/129)log2(129/129)= 0
Entropy is 0 means homogeneous. All instances are belonging to same class therefore it become leaf node in the decision tree.
With Stress = financial
Stress = financial is of occurrences 48 (8 instances are 'yes' and 40 are 'no')
Entropy (D financial) = -(8/48)log2(8/48)-(40/48)log2(40/48) = 0.650022
Gain (Dfinancial , chngGoal)= 0.322994
Gain (Dfinancial , Residence)= 0.022157
Gain (Dfinancial , FType)= 0.119715
Gain (Dfinancial , HSC_Stream)= 0.151867
Gain (Dfinancial , Infrastructure)= 0.176065
Gain (Dfinancial , Par_Curr)= 0.186707
Gain (Dfinancial , FmlyPrblm)= 0.028132
Gain (Dfinancial , CmpsEnvironment)= 0.112388
Gain (Dfinancial , AdjustPrblm)= 0.188364
Gain (Dfinancial , EnrolOthrInst) = 0.178217
ChngGoal has the highest gain; therefore it is used as decision node.

Step 16: This process goes on until all data is classified perfectly.

Step 17: Build a decision tree using above information gain



After tree construction and confusion matrix, evaluation parameters such as Recall, F-measure, Precision and Accuracy are calculated shown in Table 4.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Table 3: Confusion matrix of ID3

		Predicted class		
		No	Yes	Total
Actual class	No	182 (TP)	1 (FN)	183
	Yes	3 (FP)	34 (TN)	37
	Total			220

Table 4: Results for the ID3 decision tree algorithm using 10- fold Cross validation (Accuracy by class)

Accuracy	Class	
	No	Yes
TP rate	0.995	0.919
FP rate	0.081	0.005
Precision	0.984	0.971
Recall	0.995	0.919
F-measure	0.989	0.944
ROC area	0.957	0.957

B. Classifier Rules

IF Stress=No THEN Dropout= No
IF Stress= Financial AND chngGoal=No AND AdjustPrblm= No AND infrastructure= good THEN Dropout= No
IF Stress= Financial AND chngGoal=No AND AdjustPrblm= No AND infrastructure= V.good THEN Dropout= No
IF Stress= Financial AND chngGoal=No AND AdjustPrblm= No AND infrastructure= Poor AND Residence= Urban THEN Dropout= No
IF Stress= Financial AND chngGoal=No AND AdjustPrblm= No AND infrastructure= Poor AND Residence= Rural THEN Dropout= No
IF Stress= Financial AND chngGoal=No AND AdjustPrblm= No AND infrastructure= Excellent THEN Dropout= No
IF Stress= Financial AND chngGoal=No AND AdjustPrblm= Yes THEN Dropout=Yes
IF Stress= Financial AND chngGoal=Yes THEN Dropout=Yes
IF Stress=Other AND Residence=Urban THEN Dropout=Yes
IF Stress=Other AND Residence=Rural AND HSC_ Stream=Math THEN Dropout=No
IF Stress=Other AND Residence=Rural AND HSC_ Stream=Commerce THEN Dropout=No
IF Stress=Other AND Residence=Rural AND HSC_Stream=Arts THEN Dropout=No
IF Stress=illness AND Ftype=nuclear THEN Dropout=No
IF Stress=illness AND Ftype=Joint THEN Dropout=Yes

Use the classifier rules to improve student admission plan, tracking and help the students who have a high probability of dropping out including educational quality management planning of the university.

C. Cause of Dropout

The data collected from 220 students was analyzed to study the frequency distribution against each factor of those students who have completely decided to drop out during the course of study programme. The dropout variable has two possible values such as Yes (students who have completely decided to dropout), and No (students not interested to dropout) and based on these two groups see in Figure 2. Table 5 shows the frequency and percentage of student, a significant percentage of students (37, 16.8%) are completely decided to drop and (183, 83.2%) students will continue their study.

Table 5: frequency distribution of students in dropout

	Frequency	Percent	Cumulative Percent
No	183	83.2	83.2
Yes	37	16.8	100.0
Total	220	100.0	

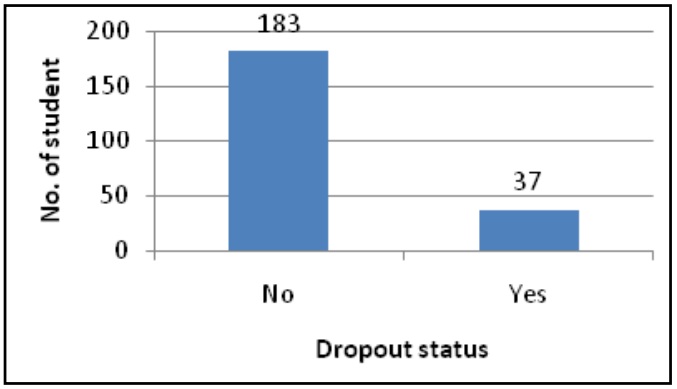


Fig. 2: Frequency Distribution of students in dropout

The reasons provided by the students for dropping out of the course (B.Tech. and BCA) are divided into four groups such as family problem, health related problem, personal problem and institutional problem were listed in Table 6. The highest dropout reasons were family reasons (8.64%), in institutional factors the highest dropout reason were campus environment (8.18%) followed by too many rules (2.73%) and low placement rate (2.27%) and in personal problem the highest dropout reason were change of goal (5.45%), adjustment problem in hostel (4.09%) and home sickness (5.45%). Whereas few students likely to dropout due to illness, home sickness, peer problem, high course fee, adjustment problems and low placement rate etc.

Table 6: Cause of dropout

Reasons	Yes	
	Number	percentage
Family Problem	19	8.64
Home sickness	12	5.45
Campus environment	18	8.18
Too many Rules	6	2.73
Low Placement rate	5	2.27
Change of personal goal	12	5.45
Adjustment problem	9	4.09
Enrolled for other institute	14	6.36

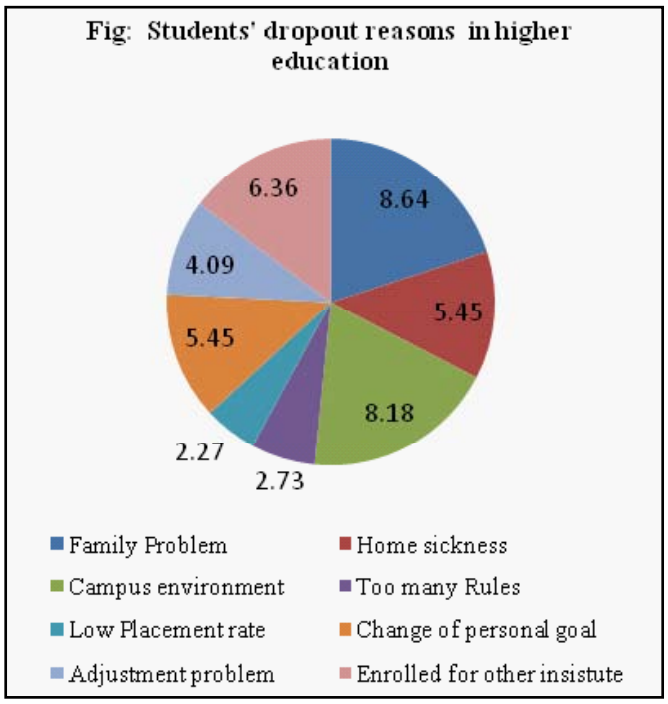


Fig.3: Student's dropout reasons at residential university represented by Pie Chart

The graphical presentation (Figure 3) also shows trend against each causing factors. The highest contribution was recorded by personal problem followed by institutional problem, family problem and health related problem.

IX. Conclusion & Future Work

This paper proposed a novel concept of correlation in attribute selection. Based on the results, conclude that a student dropout appears to be correlated with the residence, stress, family type, stream in higher secondary, satisfaction level, enrolled for other institute, change in goal, infrastructure of university, participation in extra-curricular activity, adjustment problem in hostel, and family problem.

In other word, student will dropout whose residence is urban, having some stress, who belong to joint family background, taken bio or arts as a stream in higher secondary, who are not satisfied with course, applied for other institute also, whose goal was changed, who doesn't like the infrastructure of university and having some negative attitude towards university, who does not participated in any curricular activity, who have adjustment problem in hostel and having some problem in family.

Result indicates that ID3 decision tree algorithm is best classifier with 98% accuracy. This study will also work to identify those students which needed special attention to reduce drop-out rate.

The generated knowledge will be quite useful for management of university to develop policies and strategies for better planning and implementation of educational program and infrastructure under measurable condition to increase the enrolment rate in University and to take effective decision to reduce student dropout

Future work is to study on large database of dropout student at the University using other data mining technique such as logistic regression, association and clustering in order to determine similarities and relationship between multiple factors of student who dropout.

References

[1] *M.Ramaswami and R.Bhaskaran. "A Study on Feature Selection Techniques in Educational Data Mining," Journal of computing, vol.1, no. 1, 2009.*

[2] *Han, J & Kamber, M. (2001). Data Mining: Concepts and Techniques. San Francisco (CA, USA): Morgan Kaufmann Publishers, Academic Press. 550 p. ISBN: 1-55860-489-8.*

[3] *Romera, C. and Ventura, S., " Educational Data Mining: A Survey from 1995 to 2005."Expert Systems with Applications 33, 125-146,2007..*

[4] *S. Ayesha, T. Mustafa, A.R. Sattar, and M.I.Khan, Data Mining Model for Higher Education System, European Journal of ScientificResearch, Vol.43, No.1, pp.24-29, 2010.*

[5] *D'mello, S.K., Craig, S.D., Witherspoon, A.W., McDaniel, B.T. and Graesser, A.C., "Automatic Detection of Learner's Affect from Conversational Cues." User Modeling and User-Adapted Interaction vol 18. pp. 45-80, 2008.*

[6] *Romero, C., Ventura, S., Eapejo, P.G. and Hervas, C.,." Data Mining Algorithms to Classify Students." In Proceedings of the 1st International Conference on Educational Data Mining, pp. 8-17, 2008.*