# **ID3 Classifier for Pupils' Status Prediction**

K.Nandhini, PhD. Professor Department of Computer Applications Professional Group of Institutions Coimbatore to Trichy Road K.N.Puram, Palladam[Tk]-641 662

## ABSTRACT

Predicting the pupil's status is the primary goal. Many studies have been made by a large number of scientists to explore the prediction of their research. One best solution is predicting the results based on the data source by applying some data mining techniques. This research work is to identify the prediction results by means of applying classification technique on the data source being available. There are many approaches in classification technique but this paper implements ID3 (Iterative Dichotomiser 3) Decision Tree concept which provides higher accuracy rates. This model extracts highly useful, reliable patterns from the database to ensure pupil's to achieve a higher academic output.

## **Keywords**

Classification, Decision Tree, ID3 Algorithm, WEKA Tool.

## **1. INTRODUCTION**

Data are collected in various formats like images, audios, videos, records, text files, etc. are stored in the database. These data can be efficiently extracted and mined from the repository by means of Knowledge Discovery Process often called Data Mining, so that it can be used to provide a proven solution to the problem defined. In simple words KDD can be defined as the process of extracting the useful and valid patterns from the data warehouse for resolving a problem. Data mining is being used in many areas like banking, insurance, hospital, education and in new fields of Statistics, Databases, Machine Learning, Pattern Recognition, Artificial Intelligence (AI) and Computation capabilities etc. Nowadays data mining has extended its research in the field of education.

Educational Data Mining (EDM) uses various techniques like Classification, Decision Trees, Rule Induction, Nearest Neighbors, Neural Networks, Clustering, Genetic Algorithms, Exploratory Factor Analysis and Stepwise Regression. Each technique provides different sort of results, this work computes the result by means of ID3 algorithm of decision tree. Information like Personal details, Semester percentage, Extracurricular activities, Seminar participation, Paper presentation, Certification courses etc. based on these discipline and behavior were collected from the pupil's management system, to predict the status. This paper investigates the accuracy of Decision tree techniques for predicting pupil's performance by using the Data Mining Machine Learning Tool WEKA. Performance profiling is dependent upon the motivation, attitude and marks and by the continued real-time monitoring of pupil's performance using a simple rapid response system that predicts correctly which pupil may need some extra attention in the course of their education. The model developed to achieve a measurable pupil's progress monitoring process that gives results quickly

S. Saranya M.Phil Scholar Department of Computer Science Dr. NGP Arts and Science College Coimbatore-641 048

and meets a larger educational goal benefiting stakeholders in the educational system and the wider community.

# **1.1 WEKA TOOL**

WEKA is an open source java code created by researchers at the University of Waikato in New Zealand. WEKA is a library of Java routines for various machine learning tasks. WEKA can also be used as a stand-alone learner. It provides many different machine learning algorithms, including Decision Tree (J4.8, an extension of C4.5), MLP- Multiple Layer Perceptron (a type of neural net), Naïve Bayes, Rule Induction Algorithms such as JRip, Support Vector Machine (SVM) and more.

Data mining is solely the domain of big companies and expensive software. In fact, there is a piece of software that does almost all the same things as these expensive pieces of software, called WEKA. It uses the GNU General Public License (GPL). The software is written in the Java language and contains a GUI for interacting with data files and producing visual results. It also has a general API, so you can embed WEKA, like any other library, in our own applications to such things as automated server side data mining tasks.

WEKA is preferred method for loading data is in the Attribute Relation File Format (ARFF). First define the type of data being loaded, supply the data itself. In the file, define what each column contains. In the case of the regression model, it's limited to a NUMERIC or a DATE column. Finally, supply each row of data in a Comma Delimited Format.

The main objective of this paper is to use data mining methodologies to study pupil's performance in the courses. Data mining widely used to study pupil's performance. In this paper Decision Tree Method is used for data classification. Information like personal details, percentage, extracurricular activities were collected from the department of computer science to evaluate the pupil's status. This research paper investigates the prediction of pupil's status by applying ID3 classifier technique.

## **2. LITERATURE SURVEY**

Data mining has evolved its research very well in the field of education in a massive amount. This tremendous growth is mainly because it contributes much to the educational systems to analyze and improve the growth of pupil's as well as the pattern of education.

Han and Kamber [1] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. Galit [2] gave a case study that use pupil's data to analyze their learning behavior to predict the results and to warn pupil's at risk before their final exams.

Al-Radaideh, et al [3] applied a decision tree model to predict the final grade of pupil's who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Khan [4] conducted a performance study on 400 pupil's 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 socio-economic status had relatively higher academic achievement.

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

Cortez and Silva [6] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school pupils' from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 pupils', 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 also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

Pandey and Pal [7] conducted study on the pupils' performance based by selecting 600 pupils' 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 pupils' will performer or not.

Bray [8], in his study on private tutoring and its implications, observed that the percentage of pupils' receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Srilanka. It was also observed

## **4. DATA SOURCE**

The pupil's status is determined mainly by their semester percentage, extracurricular activities, seminar presentation, certification course and paper presentation. Each pupil's has to get minimum marks to pass a semester in internal as well as that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

## 3. ID3 (Iterative Dichotomiser 3)

ID3 developed at the University of Sydney. Ross Quinlan first presented ID3 in 1975 in a book, Machine Learning, vol. 1, no. 1. ID3 is based on the Concept Learning System (CLS) Algorithm [9].

## 3.1 Splitting Criteria

A fundamental part of any algorithm that constructs a decision tree from a dataset is the method in which it selects attributes at each node of the tree.

## 3.2 Entropy

A measure used from Information Theory in the ID3 algorithm and many others used in decision tree construction is that of Entropy. Informally, the entropy of a dataset can be considered to be how disordered it is. It has been shown that entropy is related to information, in the sense that the higher the entropy, or uncertainty, of some data, then the more information is required in order to completely describe that data. In building a decision tree, we aim to decrease the entropy of the dataset until we reach leaf nodes at which point the subset that we are left with is pure, or has zero entropy and represents instances all of one class (all instances have the same value for the target attribute). We measure the entropy of a dataset, S, with respect to one attribute, in this case the target attribute, with the following calculation:

Entropy measures the amount of information in an attribute. Given a collection S of c outcomes  $Entropy(S) = S - p(I) \log 2$ p(I) where p(I) is the proportion of belonging to class I. S is over c. Log2 is log base 2.Note that S is not an attribute but the entire sample set.

If S is a collection of 14sample records with 9 YES and 5 NO sample then Entropy(S) = -(9/14) Log2 (9/14) - (5/14) Log2 (5/14) = 0.940. Notice entropy is 0 if all members of S belong to the same class (the data is perfectly classified). The range of entropy is 0 ("perfectly classified") to 1 ("totally random").

## 3.3 Gain

Gain is computed to estimate the gain produced by a split over an attribute, Gain (S, A) is information gain of example set S on attribute A is defined as

 $Gain(S, A) = Entropy(S) - S ((|S_v| / |S|) * Entropy (S_v))$ 

Where:

S is each value v of all possible values of attribute A

$$\begin{split} S_v &= \text{subset of } S \text{ for which attribute } A \text{ has value } v \\ |S_v| &= \text{number of elements in } S_v \\ |S| &= \text{number of elements in } S \end{split}$$

Gain quantifies the entropy improvement by splitting over an attribute: higher is better.

semester examination. The data set used in this study was obtained from Dr. N.G.P arts and science college, Coimbatore (Tamil Nadu) on the sampling method of computer science department of course M.Sc. (Master of Computer Science) from session 2011 to 2013. Initially size of the data is 60. By

identifying these pupils' known, we are able to decide which type of pupils' are more successful than others and provide academic help for those who are less likely to be successful.

#### **4.1 Data Preparation**

During data collection, the relevant data is gathered and the quality of data must be verified. Usually, the assembled data contains of missing or incomplete attribute, noisy (containing errors, or outlier values that deviate from expected), and inconsistent of data are common. Therefore, the collected data must be cleaned and transformed before it can be utilized in data mining system since data mining should process cleaned data in order to come out with better and or quality results. Data cleaning involves several of processes such as filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Then, the cleaned data's are transformed into a form of table that is suitable for data mining model. The cleaned data will be divided into two; training or learning data (60%) and the rest is for validating the data. These training data is applied to develop the model while the validated data is used to verify the chosen model.

### **5. RESULTS**

The data set of 60 pupils' used in this study was obtained from Dr.NGP Arts and Science College, Coimbatore (Tamil Nadu) Computer Science Department of course MSc (Master of Computer Science) from session 2011-2013.

Home Muert	Page Layout Fr	ormulas Data Review Vi	ew Add-Ins								8 - 7 3
Paste Cut -a Coty Paste Painter Cipboard	Calibri - 11 B J U	· (A' A') = = = (⊗·) (⊙· · <u>A</u> · ) = = = (≥ (⊂) (○ · <u>A</u> · ) = = = (⊂ (⊂) ()	Witap Text	General - No + No Number	- Candi Format	tional Format Cet ming - as Table - Styles - Styles	The sector	Delete Forma	Σ AutoSum * A Pill * 2 Clear * frite Editing	A Find A	
A1 .	• 🚱 🖌 Nam	ne									
A 8	C D	E F G	H 1	J K	L	M N	0	P	Q R	s	T U
1 Name Gender	Fname Fjob	Mname Mjob Nationalit	ADstreet ADcity	SSLC SSLCedu	a SSLCmed	i SSLCboarc HSC	HSCedua	a HSCmedi	HSCboard UG	UGeducat UG	mediu: UGboa
2 K.Akshaya F	B.Krishna Building	c K.Geetha Home mal Indian	Sakthi nag Coimbato	very good one	English	State boarvery good	one	English	Matriculativery good	fone En	glish Bharat
3 P.Anand M	M.K.Packi Business	P.Sivagam Home mail Indian	Vinayaga Pollachi	good one	Tamil	State boar good	one	Tamil	State boargood	one En	glish Autone
4 F.Anita ch F	J.Francis Coacher	F.Leelavat Home mail Indian	Thilakar st Coimbato	very good three	English	State boar poor	three	English	State boarvery good	three En	glish Bharat
5 F.Ann rosi F	G.Francis (Fisterman	n G.Esabel Home mal Indian	Antony sti Kanyakum	very good one	Tamil	State boargood	one	Tamil	State boarvery good	ione En	glish Bharat
6 D.Aravind M	K.Prabaka VAO	K.Andal Home mailindian	Kaveripat Krishnagir	very good one	English	State boar poor	one	English	State boargood	one En	glish Periyar
7 L.Arthi F	V.LaksmarAgricultu	r L.Jayamar Home mal Indian	Ganapath Udumalpe	poor one	Tamil	State boar poor	one	Tamil	State boar poor	one En	glish Bharati
8 R.Arul pra M	5.Raja Farmer	R.Selvi Home mal Indian	Kallukkad Namakkal	very good one	Tamil	State boargood	one	Tamil	State boargood	one En	glish Periyar
9 S.Banuprit F	B.Sampati Finance	S.Prabhav Home mal Indian	Pillaiyar k Tirupur	good three	English	State boarvery good	three	English	State boarvery good	ithree En	glish Bharat
10 K.Bharath M	P.Krishnai Farmer	K.Palaniyi Farmer Indian	Pattagapa Krishnagir	very good two	Tamil	State boargood	one	Tamil	State boarvery good	itwo En	glish Madrad
11 K.M.Bhara M	K.R.Mani Agricultu	r R.Halamm Home mal Indian	Kappachi The nilgiri	igood one	English	Matriculal poor	one	English	State boarvery good	fone En	glish Periyar
12 M.Boni m M	Mathew Farmer	Gracy Home mailindian	PayyavooiKannur	good one	Malayala	r Kerala sta very good	one	Malayalar	HS8 of exi poor	one Ma	ilayalari Kannur
13 S.Chandra M	P.Shittrari Farmer	S.Maariya Home mal Indian	Thottiyam Trichy	poor one	Tamil	State boai poor	one	Tamil	State boarpoor	one En	glish Periyar
14 J.Deepa si F	E.V.Dilee; Business	R.JayasreeGovernmeIndian	Alappuzh: Kerala	very good one	English	Kerala sta good	one	English	Kerala sta very good	ione En	glish Bharat
15 S.Divya F	T.Sreenivi Business	S.Pushpar Staff nurs-Indian	Brindavan Colmbato	very good one	English	State boar poor	three	English	State boargood	one En	glish Bharat
16 V.Elumala M	Venkatesi Farmer	Muthulak: Home mailindian	Palli patti Dharmapu	good one	Tamil	State boar very good	two	Tamil	State boargood	one En	glish Periyar
17 R.Hamsati F	<b>B.Ravi sha Business</b>	R.Shakila Home mal Indian	Sadaiyapp Tirupur	very good three	English	State boarvery good	three	English	State boarvery good	three En	glish Bharat
18 R.Harilal M	K.S.Reji Forman	K.Sulocha Home mal Indian	Marthom: Gudalur	very good one	Malayalar	r State boargood	one	English	State boargood	one En	glish Bharat
19 K.Jegan M	P.Kasi Farmer	K.Malar Farmer Indian	Sooranam Ramanath	very good one	Tamil	State boargood	one	Tamil	State boargood	one En	glish Bharat
20 K.Jayakun M	T.Krishnar Farmer	K.Poonga-Home mailindian	Purathur (Palacodu	good one	Tamil	State boar poor	one	Tamil	State boargood	one En	glish Periyar
21 T.Jeevalat F	R.Thangav Business	T.Easwari Business Indian	Thekalur Tirupur	very good one	Tamil	State boar very good	three	Tamil	State boarvery good	fone En	glish Bharat
22 S.Jeevitha F	N.Senthil Farmer	5. Tamilsel Home mal Indian	Guruvarec Bhavani	very good three	English	State boar very good	three	English	State boarvery good	three En	glish Bharat
23 K.Kalaisel F	E.Karuppa Farmer	K.Subulak Home mal Indian	Vadugapa Avinasi	very good three	Tamil	State boarvery good	three	Tamil	State boarvery good	three En	glish Bharat
24 S.Kanchar F	R.Subrams Agricultu	r S.Rajamar Home mal Indian	Pachapala S.S.Kulam	very good one	Tamil	State boargood	one	Tamil	State boarvery good	fone En	glish Bharat
25 P.Karthika F	K.Palanisa Nil	P.JayalaksFarmer Indian	Arunagirir Erode	poor three	English	State boar very good	three	English	State boargood	three En	glish Autono
H + + H saranyasss	2					14	_				

#### Fig 1: Dataset

The above fig 1 contains 60 pupils' personal details of their SSLC, HSC, UG degree percentage and family details like their parents education qualification to determine whether they are eligible to do higher studies or not and also able to determine the pupils' status. In this paper using WEKA tool to find out the eligible pupils' by using ID3 classifier.

The following step clearly specifies the ID3 classifier performance in WEKA tool.

**Step-1:** first we have to import the data's into the WEKA explorer and then preprocess the data. The following screen identifies the preprocessing step.

Preprocess Classify Cluster Asso	ociate Select attributes Visual	20					
Open file	Open URL	Open DB	Gene	rate	Undo	Edit	Save
Filter Choose None Current relation				Selected attribute			Apply
Instances: 50	At	tributes: 29		Missing: 0 (0%)	Distinct: 3	Uniq	ue: 0 (0%)
Attributes	None	Invert	Pattern	No. Label 1 very good 2 good	4	Count 26 15	
No.         Name           8         ADstreet           9         ADstreet           10         \$SLC           11         \$SLCeducationtype           12         \$SLCeducationtype           13         \$SLCoordofstudy           14         HSC	а У		-			12	
16 HSCmediumofstudy	,			Class: Parttimejob (Nom)			✓ Visualize All
17 HSCboardofstudy 19 UiGebcation type 20 UiGebcation type 21 Uicbeardiniversity 22 UiCaclegename 23 Collegedcation 24 Seminarparticipatio 26 Paperpresentation 27 Numberofcertificati 28 Attendedaryplacer 29 Partimetob	ofstudy n oncourses nent			20			
	Remove						
Status OK							Log 💉 0

Fig 2: Preprocessing

Step-2: next classify the data by entering the cross validation

value as 10. The screen will be like this,

Weka Explorer								- 0 - X
Preprocess Clessify Guster Associate	Select attributes Visualize							
Classifier								
Choose Id3								
Test options	Classifier output							
Use training set								
Suppled test set	Nume = K.P.Vineetha: no Nume = S.Vinoth: yes							-
Cross-validation Folds 8								
🕐 Percentage split 🛛 🛸 🚳	Time taken to build model: 0.02 se	conds						
More options	=== Stratified cross-validation == === Summary ===	-						
(Non) Parttinejob	Correctly Classified Instances	a		0	<i>v</i> .			
<b>0 1 1 1 1 1 1</b>	Incorrectly Classified Instances	0		0	*			
Start	Kappa statistic	1						
Result list (right-click for options)	Mean absolute error	NeW						
10-29-58 - trees 1d3	Root mean squared error Relative sheelute error	No.W						
10139130 0000000	Root relative squared error	NoN						
	UnClassified Instances	50		100	\$			
	Total Number of Instances	.50						
	Detailed Accuracy By Class							
	TP Bate FP Bate	Precision	Recall	F-Measure	RDC Area	Class		
	0 0	0	0	0	0.5	no		
	0 0 Weighted Avg. NaN NaN	0 NeN	0 NoN	0 NeN	0.5 NeW	yes		
	Confusion Batrix							
	a b < classified as							E
	0 0 1 a = no							
	0 0   b = yes							
								Ŧ
Status								
OK								Log X × 0

## Fig 3: Classification

Step-3: third step selects the attributes which gives the first

priority to derive the tree whether as a left node or right node.

Weka Explorer		
Preprocess Classify Cluster Associate	ielect attributes Visualize	
Attribute Evaluator		
channel Terfort attraction to the Toront		
Crocse InfoGanvcchoucetval		
Search Nethod		
diana and a stranger		
Choose  Kanker -1 -1.79/69313486	315/6308-14-1	
Attribute Selection Mode	Attribute selection output	
Use full training set	0.640077 5 Mname	
Consultation Edite 10	0.459686 22 UEcollegename	-
Clos-raioauni 1005 10	0.362911 9 ADcity	
Seed 1	0.193729 21 UGboardUniversityofstudy	
-	0.17451 4 Fjob	
(Non) Parttinejob 🔹	0.139507 2 Gender	
	0.128391 18 UG	
Start Stop	0.12382 10 SSLC	
Result list (right-click for options)	0.072184 19 UGeducationtype	
	0.05445 25 Paperpresentation	
10:42:14 - Ranker + InfoGainAttributeE	0.052816 27 Numberofcertificationcourses	
	0.04876 11 SSLCeducationtype	
	0.036976 15 HSCeducationtype	
	0.036963 6 Hjob	
	0.02889 17 HStboardofstudy	
	0.01091 12 SSLCmediumofstudy	
	0.010437 25 Sportsportsporticipation	
	0.009546 16 Holdedradorstudy	
	0.000321 20 Attendedanyplatement	
	D 005269 13 931/haardofetude	
	0.00579 20 IIBerdiumofstady	
	0.001484 24 Seminarparticipation	
	0.000804 23 Collegelocation	e
	0 7 Nationality	
	Selected attributes: 1,3,8,5,22,9,21,4,2,18,10,19,25,27,11,15,6,17,12,26,16,28,14,13,20,24,23,7 : 28	
4 III →		v
Status		100 000
ok		

#### Fig 4: Select Attribute



#### Step-4: The following diagram specifies the overall data set

visualization.

#### Fig 5: Visualization

# 6. CONCLUSION AND FUTURE

# ENHANCEMENT

This paper analogous classification techniques which is ID3 classifier by using the data mining tool WEKA. The ID3 classifier used to predicting the right pupil's who are eligible to do their PG courses. The efficiency of the ID3 algorithm is highly accurate than the other mentioned algorithms. This paper analyses which pupil performing good or bad in their upcoming courses. The above steps denote the ID3 performance using WEKA. In future the ID3 algorithm combined with genetic algorithm to predicting the pupil's status.

# 7. REFERENCES

- [1] Andrew Colin, "Building Decision Trees with the ID3 Algorithm", by: Dr. Dobbs Journal, June 1996.
- [2] Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.
- [3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

- [4] M. Bray, The Shadow Education System: Private Tutoring and Its Implications for Planners, (2<sup>nd</sup> ed.), UNESCO, PARIS, FRANCE, 2007.
- [5] 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.
- [6] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.
- [7] S. T. Hijazi, and R. S. M. M. Naqvi, "Factors Affecting Student's Performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.
- [8] U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.
- [9] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp.84-87.