Analysis of Students' Performance Evaluation using Classification Techniques

V. Shanmugarajeshwari
III MCA
Ayya Nadar Janaki Ammal College
Sivakasi
TamilNadu, India
sanmugaraj93@gmail.com

R. Lawrance
Director, Department of MCA
Ayya Nadar Janaki Ammal College
Sivakasi
TamilNadu, India
lawrancer@yahoo.com

***Abstract*—Educational Data Mining is one of the emerging disciplines which includes the process of analyzing the students' details using different attributes. The attributes such as students' name, roll number, previous semester marks, attendance, assignment, seminar performance, lab work and gender are used to evaluate the students' performance (Pass / Reappear). In this paper, classification techniques are described and used for educational data mining. The classification process is based on C5.0 algorithm with good classification accuracy. The system is helpful to the learners as well as to the teachers for the academic performance evaluation. It is a warning system for the students' to improve their study performance.**

Keywords—Educational Data Mining (EDM), Feature Selection techniques, Classification, Students' Performance.

I. INTRODUCTION

Data mining is used to extract the meaningful information from large data using some patterns. It has been used in many applications such as educational data mining, web mining and text mining. Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students' and the settings in which they learn [1]. These methods are different from the standard data mining methods. There are three main goals in educational data mining, [2]

Pedagogical: To help in the design of didactic contents and improve the academic performance of the students.
Managerial: To optimize the organization and maintenance of educational infrastructures, areas of interest and do research.
Commercial: To help in students' recruitment

In the present work, educational data mining techniques are described. The data mining applications and classification techniques can be used to develop the education sector.

II. LITERATURE REVIEW

Literature survey refers to a critical summary. Literature reviews contextualize research about a topic. A literature review is an evaluative report of studies found in the literature related to selected area. The review should describe, summarize, evaluate and clarify this literature. It should give a theoretical basis for the research and help you determine the nature of our own research. [3] It Reviews that what have already been done in the framework of a topic. Therefore, on the basis of the existing knowledge, everyone can build up innovative idea and concept for further research purpose. [4] The major benefits of literature survey are,

Assessment of the current state of research on a topic. Identification of the experts on a particular topic. Provide a context for the research. Identification of key questions about a topic that need further research. Determination of methodologies used in past studies of the same or similar topics. [5]

Feature Selection

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. [6] Several feature selection methods are available in data mining for selecting the relevant attributes.

Baradwaj, et al. [7] implemented an ID3 decision tree learning algorithm with the help of the previous example which included the training set. In their study, they described the reasons for using decision tree algorithms. They calculated the entropy values for each attribute in the data set and then calculated information gain for each attribute. Then they selected the root node as an attribute which has the highest information gain and found which attribute was the next decision node until they ran out of attributes. Finally, their ID3 classification algorithm has generated the decision tree for weather data set.

Adhatrao, et al. [8] implemented an ID3 and C4.5 Decision tree technique on educational data mining. First-year students' data was collected. Their system was used to predict the result of same students in second year. The accuracy of the ID3 algorithm is 75.145% and that of C4.5 is 75.145%. They concluded study can be used to predict the students' result based on previous semester marks.

978-1-4673-8437-7/16/$31.00 ©2016 IEEE

Z. J. Kovacic *et al.* [9] presented a case study on educational data mining to identify up to what extent enrollment data can be used to predict students' success. They used CART and CHAID decision trees and the accuracy of classifiers obtained was 59.4 and 60.5 respectively.

Ramaswamy *et al.* [10] designed a technique on students' data which has 33 features including class label. 6 feature selection techniques were applied on the data set, for selecting the relevant attributes. These attributes have a relevant value to the students result. They used different classification algorithms. Voted perception showed the highest predictive accuracy of 89%.

Acharya *et al.* [11] applied feature selection techniques and data mining algorithms on students' data. The data have been collected from St Xavier's College, Kolkata. Different feature selection techniques were applied on the data for extracting the relevant attributes and discarding irrelevant attributes. They got 79% accuracy.

III. DATA SET DESCRIPTION

A. Data Collection

The input data have been collected from Ayya Nadar Janaki Ammal College, Sivakasi, TamilNadu, India from the students of computer Applications department (Master of Computer Applications). Initially the data size is 47 records X 12 attributes.

B. Data Preparation

The students' data are collected from a survey. Mark details, personal information as well as family background details are collected. The attributes are MED (Medium of Study), FG (First Graduate), RESI (Residence), LIVLOC (Living Location), FSIZE (Family Size), FEDU (Fathers Qualification), MEDU (Mothers Qualification), FINC (Family Annual Income), ATT (Attendance), HSCM (Higher Secondary Marks) and PSM (Previous Semester Marks). RESULT is the dependent variable. Attributes used in this study are described in Table I.

TABLE I. DATA SET INFORMATION

Attributes	Description	Possible Values
MED	Medium of study	Tamil, English
FG	First Graduate	Yes, No
RESI	Residence	DayScholar, Hostel
LIVLOC	Living Location	Urban, Rural
FSIZE	Family Size	2,3,4, >4
FEDU	Fathers Qualification	PG, UG, HSC, SSLC, Others, Nil
MEDU	Mothers Qualification	PG, UG, HSC, SSLC, Others, Nil
FINC	Family Annual Income	High, Medium, Low
ATT	Attendance	Good, Average, Poor
HSCM	Higher Secondary Marks	Good, Average, Poor
PSM	Previous Semester Marks	Grade A, Grade B, Grade C, Grade U
RESULT	Performance in Last Semester	Pass, Reappear

(FINC - Family Annual Income): It will be categorized into 3 classes: High - above 100000 Rs, Medium- 30000 Rs to 99999 Rs, Low – below 29999 Rs per month.

(ATT - Students Attendance): It is split into 3 classes, that is, Good – above 95, Average 80 – 94, Poor below 80.

(HSCM - Higher Secondary Marks): Based on their marks, it will be categorized into 3 groups, such as Good –Above 60, Average 40 – 59, Poor – below 40.

(PSM - Previous Semester Marks/Grade): It is obtained from MCA course. It is split into four class values: Grade A - Above 80, Grade B – 65-79, Grade C – 51-64, Grade U - Below 50.

Here, all variables are in the form of categorical values. The Result is the dependent variable and all other variables are the predictor variables.

C. Preprocessing

The students' data have been taken for preprocessing step. During preprocessing all individual tables are combined into a single table formed with all sufficient data and also missing values are also omitted. Later FINC, ATT, HSC, and PSM were classified using If Then Rules. Preprocessing can increase the classification accuracy. The proposed system of Pre-processing contains two steps.

Step 1: Removing Missing Values

Step2: Categorization

TABLE II. SAMPLE DATA SET-1

Attributes	Description	Possible Values
MED	Medium of study	Tamil, English
FG	First Graduate	Yes, No
RESI	Residence	DayScholar, Hostel
LIVLOC	Living Location	Urban, Rural
FSIZE	Family Size	2,3,4, >4
FEDU	Fathers Qualification	PG, UG, HSC, SSLC, Others, Nil
MEDU	Mothers Qualification	PG, UG, HSC, SSLC, Others, Nil
FINC	Family Annual Income	High, Medium, Low
ATT	Attendance	Good, Average, Poor
HSCM	Higher Secondary Marks	Good, Average, Poor
PSM	Previous Semester Marks	Grade A, Grade B, Grade C, Grade U

RESULT	Performance in Last Semester	Pass, Reappear

TABLE III. SAMPLE DATA SET-1.1

MEDU	FINC	ATT	HSC	PSM	Result
No	Medium	Good	Good	Grade A	Pass
No	Low	Good	Good	Grade B	Pass
No	Low	Good	Good	Grade B	Pass
UG	High	Good	Good	Grade A	Pass
Others	Medium	Good	Good	Grade A	Pass
No	Medium	Average	Average	Grade U	Reappear
SSLC	Low	Average	Good	Grade B	Pass
Others	Medium	Average	Good	Grade U	Reappear
No	Medium	Average	Good	Grade A	Pass
No	High	Average	Good	Grade U	Reappear

Based on the Pass and Reappear category the result will be predicted. After eliminating the missing values and errors, the size of the data is 44 records with 12 attributes. Sample data set 1 and 1.1 are described in Table II and Table III.

IV. METHODOLOGY

Now-a-days, educational data mining plays a major role in the field of education. In the present work educational data mining is used to improve the students' performance using feature selection and classification techniques.

A. Feature Selection

For feature selection, a number of techniques are available in data mining some of them are Chi-Squared Feature selection (CFS), Information gain feature selection (IGFS), Gain Ratio Feature Selection (GRFS) and Correlation based feature selection (CBFS) . Feature selection is also known as attribute selection [12]. For selecting the relevant attributes and discarding the irrelevant attributes, various feature selection techniques were applied on the preprocessed data set which has 44 samples. We concentrated on pick out the top attributes.

B. Gain Ratio Feature Selecction

One of the best feature selection method is Gain Ratio Feature Selection. The purpose of feature selection method is to extract the relevant features and discard the irrelevant features[13]. In Gain Ratio Feature selection, the subset has been selected using Entropy D_j value (1) and Information Gain $_A(D)$ (3). Gain ratio (4) is calculated using the equations (1) and(3)

$$Entropy\left(D_j\right) = -\sum_{j=1}^{m} p_j \log_2\left(p_j\right) \quad (1)$$

$$Info\ Gain(D,A) = Entropy\left(D_j\right) - \sum_{j=1}^{v} \frac{D_j}{D} * Entopy\left(D_j\right) \quad (2)$$

$$Gain\ Ratio(A) = Entropy(D) - Information\ Gain_A(D) \quad (3)$$

$$InformationGain_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Entropy(D_j) \quad (4)$$

From the above equations, Gain Ratio subset selector has been produced. It is used to improve the classification accuracy because it extracts the relevant attributes only. Here, redundant attributes are removed .

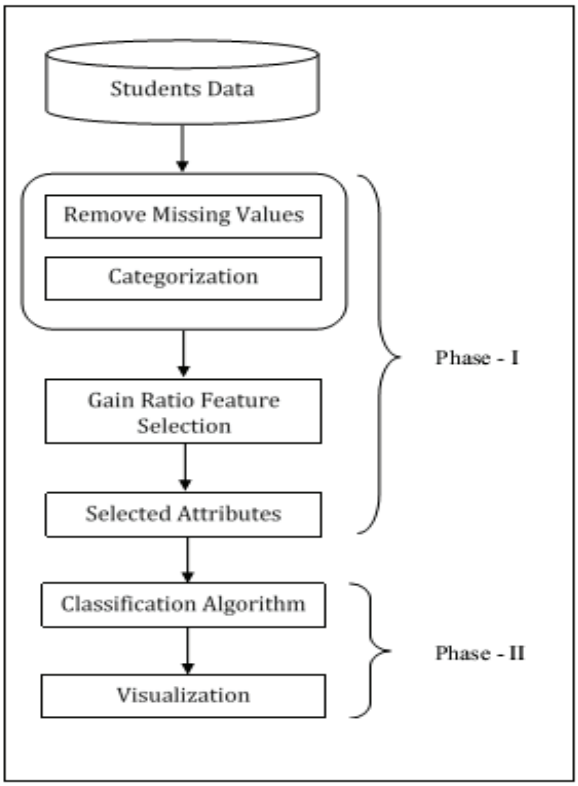


Fig.1. Methodology Framework

Figure 1 Methodology Framework using two phases: The removing missing values are removed in the preprocessing phase. The attributes are transformed into categorized format using categorization process phase.

V. CLASSIFICATION

Classification is the process to classify the data objects with the predetermined class labels. Classification is one of the supervised machine learning algorithms [14]. In classification, the preprocessed dataset has been divided into two sets, such as training set and test set. Using the training set the classifier model has been developed. The Test set was applied on the classification model, and it will be classified.

A. Decison Tree

A decision tree is a simple structure where each non-terminal node represents a test or decision on the considered data item. A decision tree can be used to classify an instance by starting at the root of the tree and moving through it until a leaf node [15].

Build Tree (Node t, Training Database D, Split Selection Method S)

1. Apply S to D to find Splitting criterion
2. If (t is not a leaf node)
3. Create Children nodes of t
4. Partition D into children Partitions
5. Recurse on each partition

End if

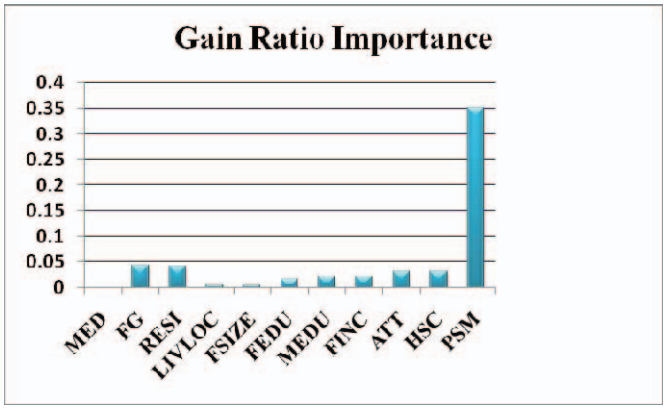
VI. EXPERIMENTAL RESULTS

We have taken the students data from Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India. During preprocessing individual tables were combined together and errors were removed. Feature selection techniques have been applied on the preprocessing data set, and relevant attributes only selected. Here, Gain ratio feature selection is applied on the data set and the result was shown in table V and table VI.

TABLE IV. SELECTED ATTRIBUTE

Method	Selected Attributes				
Gain Ratio	PSM	FG	FEDU	MEDU	FINC

Fig.2. Gain Ratio Importance



The data set of 47 students is used in this study Students' Mark details, Personal information and Family Background have been collected through a survey. It has some missed values, noisy data, irrelevant data, unknown data and classification technique is used to preprocess data. In pre-processing, missing values are omitted and FINC, ATT, HSC and PSM are categorized using If-Then Rules. Independent variable, Previous Semester Marks (PSM) external and internal marks of each subject. The Internal mark and external mark should be above 23, if it is true the candidate comes under Pass Category, or else the candidate comes under Reappear Category. After removing the missing values all the tables (Mark details, Personal information, Family Background) are merged together. Now the size of the data is 44.

TABLE V. SELECTED ATTRIBUTE

Methods	Attributes					
Chi-Squared	MEDU	RESI	FG	FEDU	ATT	MED
Information Gain	MEDU	FEDU	FINC	ATT	RESI	FG
Gain Ratio	HSC	RESI	MEDU	FG	FINC	FEDU

Feature selection methods have been applied on the data set, for selecting relevant attributes. In Table V, three feature selection methods and the results are described. Using R language feature selection methods are applied on the data set, and relevant attributes are selected. From Table V, it is revealed that Gain Ratio Feature Selection Method is the best Feature selection method. The selected attributes are HSC, RESI, MEDU, FG, FINC and FEDU. The outcome of the feature selection would be a rank list of predictor's dependent variable and the result of the feature selection are presented in a bar chart [16].These 6 are the top predictor attribute which give the best result. Gain Ratio Importance is displayed Figure 2.

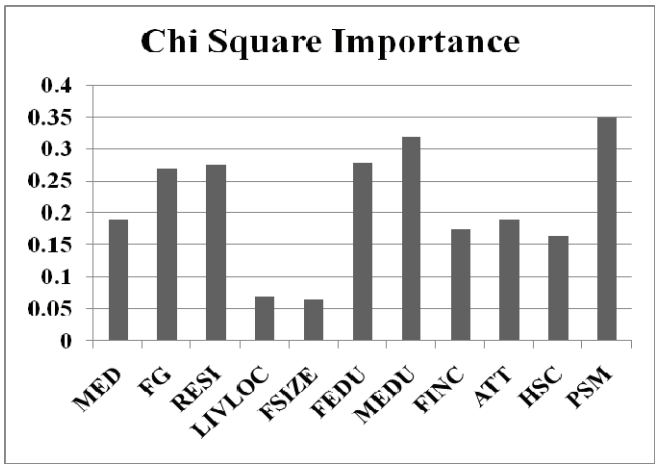


Fig.3. Chi-Square Importance (PSM is the Dependent Variable)

One of the popular feature selection method is chi-square (X^2). It is a statistical test that can be used to determine whether observed frequencies are significantly different from

expected frequencies or not, which is used to select the predictor variable. Based on observed frequency and expected frequency Chi-Square filter has been calculated. Based on the Information Gain importance the attributes were selected. Which is shown in Figure 3. Used to select the relevant attributes. The selected attributes are described in Table VI. That is MEDU, RESI, FG, FEDU, ATT and MED.

The feature selection using Chi-squared method takes very long time (0.20 Milliseconds), Information gain method takes (0.14 Milliseconds), and Gain Ratio method takes very less time (0.08 Milliseconds) Shown in Figure 4

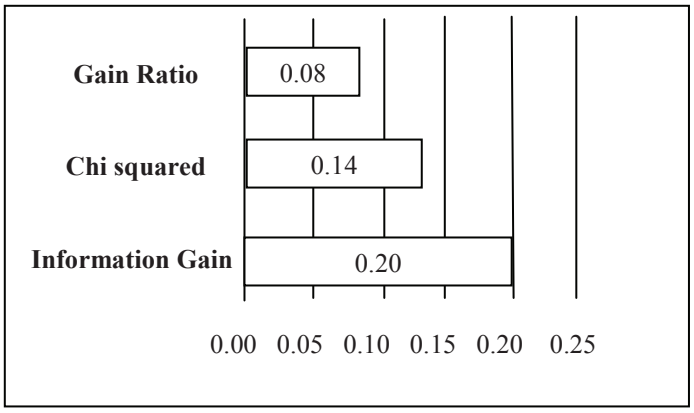


Fig.4. Time Complexity of three Feature Selection Methods

Time Complexity (in milli sec)

When compared to Chi-Squared and Information gain methods, Gain Ratio is the best method which selects the best attributes. The result of the comparative study of three different feature selection method is described in Table V.

Correctly Classified Instances are 40, which contributes 97.72 % classification accuracy. Grade A, Grade B, Grade C , Grade U are the class labels. Students comes under this category.

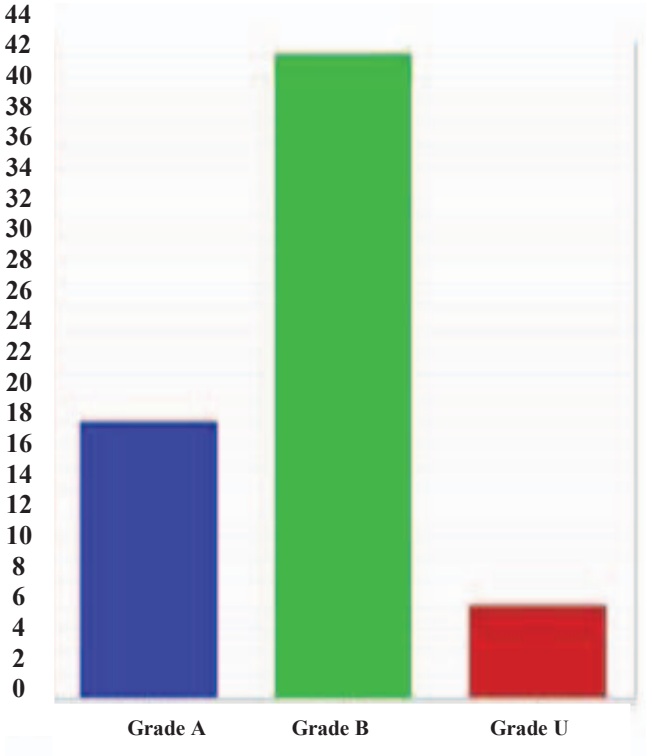
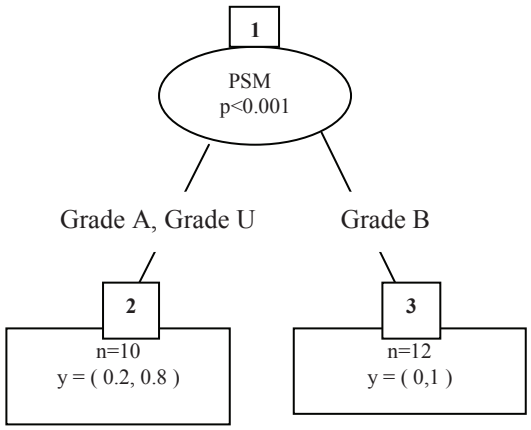


Fig.5. Histogram for Students Grade

The selected attributes have been applied on the classification technique. Decision tree classification was used here to predict the student performance. Entropy (1) and information gain (3) have been calculated. Based on the entropy value and information gain value, the gain ratio was calculated. Here according to the Gain Ratio, PSM – Previous Semester Mark is the root node of the tree. PSM has 3 categories such as Grade A, Grade B and Grade C. Based on the selected attributes this step will be recursively applied on all attributes, until there is no node to split. The decision tree has been built, and is shown in figure 6.

Fig.6. Result of Decision Tree



Training data is applied on the data set and the classifier model has been developed. The accuracy assessment should be an important part of any classification. It testifies whether the previously presented classification algorithms are either "Right" or "Wrong". The proposed system of classification accuracy is measured by values Shown in the table VI and VII,

TABLE VI. CONFUSION MATRIX

	True Reappear	True Pass
Predicted Reappear	4	0
Predicted Pass	0	40

$$\text{Accuracy} = \left(\frac{TP + TN}{TP + FN + FP + TN}\right) \quad (5)$$

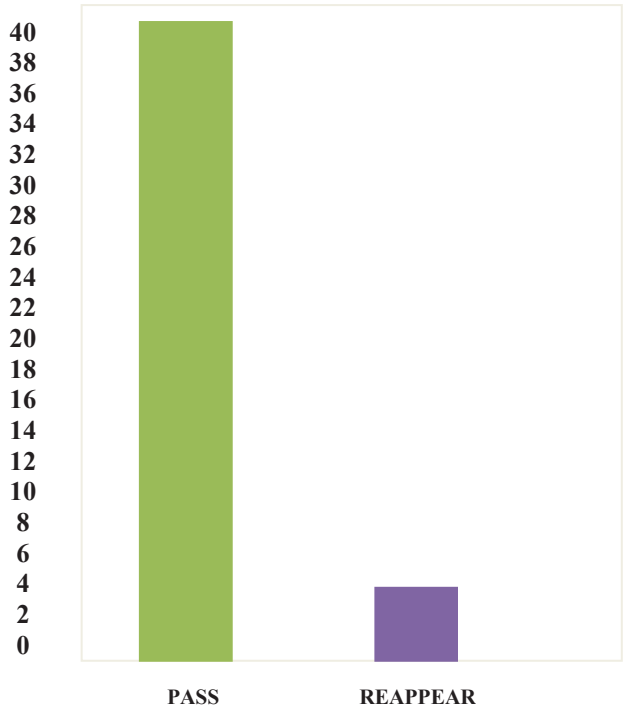
$$\text{Accuracy} = \left(\frac{4+40}{4+0+0+40}\right) = 100$$

TABLE VII. CLASSIFICATION ACCURACY

	True Reappear	True Pass
Predicted Reappear	4	0
Predicted Pass	0	40
Classification Accuracy %		100

Based on the confusion matrix in Table VI, the accuracy has been calculated [17]. The Accuracy obtained for this model is 100 percent. Pass and Reappear are the class labels. Students come under the category of Pass and Reappear shown in the figure 7.

Fig.7. Classified data



VII. CONCLUSION

Classification techniques are used to make better decisions in the educational data mining. In the proposed work students' performance has been predicted using decision tree classification methods. At the end of phase I, missing values were removed and categorized. Feature selection technique is applied on preprocessed data to select the relevant attributes. PSM, FG, RESI, HSC, MEDU, and FINC are the selected attributes. These attributes are fed into phase II. At the end of phase II, the classified students result such as PASS and REAPPEAR are obtained. The C5.0 classification algorithm has 100% classification accuracy. The proposed algorithm is compared with bench mark algorithms such as Decision tree induction and Naïve Bayes. This study is a beneficial to the teachers and the students. The system is also helpful to find the students, who need special consideration. In future, this work can be implemented in cloud computing in order to obtain more security for the heterogeneous dataset [18].

VIII. REFERENCES

[1] Bakar, R.S., and Yacef, K, "The state of educational data mining in 2009: A review and future visions." JEDM-Journal of Educational Data Mining 1.1 (2009):3-17

[2] Barracosa, J.I.M.S.2011. Mining Behaviors from Educational Data .

[3] http://library.queensu.ca/webedu/grad/Purpose_of_the_Literature_Review.pdf

[4] http://ar.cetl.hku.hk/am_literature_reviews.htm

[5] http://libguides.unf.edu/content.php?pid=496677andsid=4082503

[6] http://en.wikipedia.org/wiki/Feature_selection

[7] Baradwaj, B. K., and Pal, S. 2012. Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.

[8] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., and Honrao, V, "Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms." arXiv preprint arXiv, Vol.3, No.5, September 2013, pp.39-52

[9] Kovacic, Z. "Early prediction of student success: Mining students' enrolment data."

[10] Ramaswami, M., and Bhaskaran, R. 2009. A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924.

[11] Acharya, A., & Mukherjee, S. "Modeling Value Chain Analysis of Distance Education using UML." International Conference on Modeling, Optimization, and Computing. Vol.1298. No. 1. AIP Publishing, 2012.

[12] Zhao, Y. "R and data mining: Examples and case studies." Academic Press, 2012.

[13] Karegowda, A. G., Manjunath, A. S., and Jayaram, M. A. "Comparative study of attribute selection using gain ratio and correlation based feature selection." International Journal of Information Technology and Knowledge Management 2.2, pp.271-277.

[14] http://www.saedsayad.com/classification.htm

[15] Quinlan, J. R. "Induction of decision trees." Machine learning 1.1 1986, pp. 81-106

[16] Huang, Huaming., "Introduce Data mining with Rapid Miner" Syracuse University, EECS 2008

[17] www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

[18] Kumar, KJ Latesh, and R. Lawrance. "Novel Approach: Deduplication for Backup Systems Using Data Block Size." Computational Intelligence in Data Mining-Volume 1. Springer India, 2015. 365-373.