Predicting Students' Performance using Modified ID3 Algorithm

Ramanathan L[1], Saksham Dhanda[2], Suresh Kumar D[3]
[1]Assistant Professor (Senior), SCSE, VIT University, Vellore-632014, Tamil Nadu, INDIA
[2,3]School of Computing Science and Engineering, VIT University, Vellore-632014, Tamil Nadu, INDIA
[1]lramanathan@vit.ac.in
[2]saksham.dhanda@gmail.com
[3]sureshd_1990@yahoo.in

***Abstract*—The ability to predict performance of students is very crucial in our present education system. We can use data mining concepts for this purpose. ID3 algorithm is one of the famous algorithms present today to generate decision trees. But this algorithm has a shortcoming that it is inclined to attributes with many values. So , this research aims to overcome this shortcoming of the algorithm by using gain ratio(instead of information gain) as well as by giving weights to each attribute at every decision making point. Several other algorithms like J48 and Naive Bayes classification algorithm are also applied on the dataset. The WEKA tool was used for the analysis of J48 and Naive Bayes algorithms. The results are compared and presented. The dataset used in our study is taken from the School of Computing Sciences and Engineering (SCSE), VIT University.**

Keyword- data mining, educational data mining (EDM), decision tree, gain ratio, weighted ID3

I. INTRODUCTION

Educational Data Mining (EDM) is attracting a lot of researchers for developing methods from educational institutions' data that can be used in the improvement of quality of higher education. EDM uses the large amounts of information present in the educational institutes' databases about teaching-learning practices for the development of models which are beneficial for all the participants in the educational process.

The prediction of students' performance in any higher institute has become one of the most important needs of that institute in order to improve the quality of the teaching process of that institution. Through this process we get to know the needs of the students and hence we can fulfil those needs to get better results. Students who need special attention from the teachers can also be identified through this process. A number of algorithms are available for predicting the performance of students. Some of them being Artificial Neural Network (ANN), Decision Trees, Clustering, Naive Bayes algorithm, Decision Trees being most commonly used.

II. RELATED WORK

Bhardwaj and Pal [1] conducted a performance analysis on 50 students whose records were taken from VBS Purvanchal University, Jaunpur (Uttar Pradesh) with the objective to study student's performance using 8 attributes. Decision tree method was used to classify the data. Study helped teachers to improve the result of the student.

Srimani and Kamath[2] used various methodologies to study the application of various data mining algorithms for the performance analysis of the learning model. Various classification models like Bayesian, Multilayer perceptron(MLP),Decision Tree using j48,Rule Based RIPPER were used to implemented on dataset for class 1 to 7 that consist of 3500 data instances and 99 attributes to predict progress of each child. Results obtained from all algorithms found to be accurate and hence the model is justified with accuracy (almost 100).

Yadav, Brijesh and Pal[3]conducted study to predict students performance with 48students dataset and 7 attributes obtained from VBS Purvachal University,Janupur (UP),India on the sampling method of computer Applications department of course MCA(master of Computer Applications) from session 2008 to 2011.Different Decision tree algorithms like ID3,C4.5,CART were used for classification. Results show that CART is the best algorithm for classification of data. This study will help teachers to identify those students who need special attention and also this work will help to reduce fail ratio.

C.Marquez-vera and Venturia[4] used data mining techniques to predict school failure with 10 attributes and 670 middle-school students data from Zacatecas, Mexico classifying unbalanced data by rebalancing data and cost-sensitive classification approaches were used along with 10 classification algorithms and 10 fold-cross validation to compare and select best approach to predict students who might fail cost-sensitive classification approach is considered as best approach for classification.

Sembiring, Zarlis et al [5] aimed to apply the kernel method as data mining techniques to analyze the relationships between students' behaviour and their success by using Smooth Support Vector Machine (SSVM).

Nandini and Saranya [6] attempted to extract useful, reliable patterns for predicting pupil's performance by applying ID3 algorithm on the database of Dr. NGP Arts and Science College, Coimbatore and found out that efficiency of ID3 algorithm is good.

Kabra and Bichkar [7] discussed the use of decision trees in predicting the performance of students by using the data of 346 students of SGR Education Foundation's College of Engineering and Management and results showed that the generated tree is only 60.46% accurate, that is, only 209 instances out of 346 were correctly classified.

Pandey and Sharma [8] applied many algorithms like J48, NBtree, Reptree and Simple cart on the dataset collected from Manav Rachna College of Engineering, Faridabad for predicting the performance of students and found that J48 decision tree algorithm was best suitable for model construction.

III. ID3 ALGORITHM INTRODUCTION

ID3 algorithm is a mathematical algorithm used for generating decision tree from a dataset. This algorithm was invented by J. Ross Qinlan in 1979.

A. Slitting Criteria

The criteria used for choosing an attribute at each node of the tree is called splitting criteria. Splitting criteria used in ID3 is Entropy. Entropy is the measure of randomness in a dataset. So, higher the entropy, more is the information required to describe the data.

Given a set of data S with m outcomes, Entropy(S)= S-p(I)log2p(I), where p(I) is the proportion of dataset S belonging to class I. We tend to decrease the entropy of dataset until we reach leaf nodes. At this point, the entropy of the dataset is least, that is, 0. In order to decrease the entropy, entropy of each attribute is calculated initially and the attribute with least entropy is selected as root of the tree.

For choosing attributes at lower levels of the tree, we need to find information gain at each step. Gain is calculated to check the amount of information gained by a split over an attribute. Attribute with maximum information gain is selected as a node of the tree in further steps.

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

Where

S_v = set of values of S for which attribute A belong to class v

$|S_v|$ = number of elements in S_v

B. Drawbacks:

One of the main drawbacks of the ID3 algorithm is that it is inclined towards the attributes with more values, that is, it tends to select that attribute as a node which has more values. This can be a wrong selection and hence, as a result, the tree generated is not very much efficient.

IV. PROPOSED ALGORITHM (WEIGHTED ID3)

For removing the inclination of traditional ID3 algorithm towards attributes with many values, an improved weighted ID3 (wID3) algorithm is proposed in this paper. In this, the attribute with highest Gain Ratio (not information gain) is multiplied with a weight which gives it a new value, and among the new values, attribute with highest Gain Ratio is selected as a node of the tree. Also, information gain is replaced by gain ratio, which in more normalized. This whole process overcomes the inclination problem.

Gain Ratio (A) = Gain(S,A) / Entropy(S)

For calculating the weights, we check the relevance between condition attributes and decision attributes by using a correlation function. The value between the condition attribute A and the decision attribute D is given by:

$$AF_D(A) = (\sum^{v}_{i=1} |A_{i1}| - |A_{i2}|) / v$$

Where v is the number of different values A is having.

Now, the modified weight of condition attribute A will be

$w_A' = AF_D(A) / (\sum^{n}_{j=1} AF_D(j)$, if the gain ratio of A is max

otherwise

$w_A' = 1.$

So, the modified gain ratio of condition attribute A will be

Gain Ratio'(A) = w_A' x Gain Ratio(A)

Here we can notice that the weight of the attribute with initial maximum gain ratio is modified and rest have the same gain ratio as earlier. In this way, gain ratio of attributes satisfying the condition is reduced and hence , it overcomes the inclination problem.

V. DATASET

The Data set for the study has been collected from CSE branch of VIT University. This data set consists 304 instances and each instance consists of 10 attributes. The study considers the academic performance of the students in CSE branch in several subjects in the winter semester of the year 2011. Initially data is collected in an excel sheet and initial preprocessing is done manually by filling the missing data values by standard data and various inconsistencies has been removed. Some irrelevant attributes have been removed manually to maintain the quality of the classifier. Table 1 shows the attributes description and their possible values.

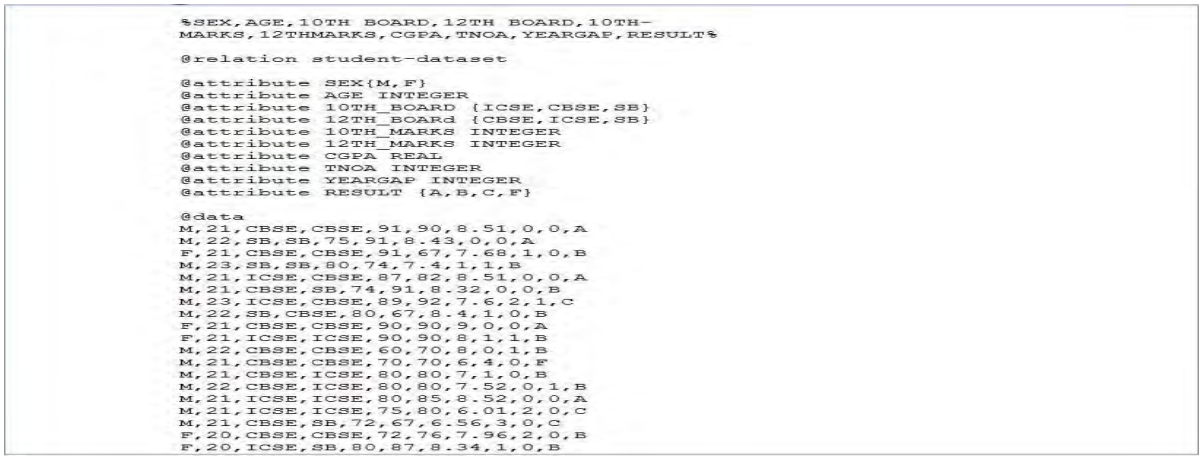
ATTRIBUTES TABLE

S.N.	*Name*	*Description*
1	*TNOA*	*Total number of arrears*
2	*CGPA*	*Cumulative grade point*
3	*AGE*	*AGE*
4	*SEX*	*SEX*
5	*10^{th} Board*	*CBSE,ICSE,SB*
6	*12^{th} Board*	*CBSE,ICSE,SB*
7	*10^{th} Marks*	*Marks obtained in 10^{TH}*
8	*12^{th} Marks*	*Marks obtained in 12^{TH}*
9	*YEAR GAP*	*Gap b/w 12^{TH} and 1^{ST} Year*
10	*RESULT*	*PASS, FAIL*

TABLE1

VI. IMPLEMANTATION METHODOLOGY

Two out of three algorithms used in this research, i.e. J48 and Naïve Bayes algorithm, are implemented using the WEKA tool. WEKA is an open source java code created by researchers at University of Walkato in New Zealand. It provides many different machine learning algorithms, including Decision Tree, MLP, Naïve Bayes, Support Vector Machine and more.



```
%SEX,AGE,10TH BOARD,12TH BOARD,10TH-
MARKS,12THMARKS,CGPA,TNOA,YEARGAP,RESULT%

@relation student-dataset

@attribute SEX{M,F}
@attribute AGE INTEGER
@attribute 10TH_BOARD {ICSE,CBSE,SB}
@attribute 12TH_BOARd {CBSE,ICSE,SB}
@attribute 10TH_MARKS INTEGER
@attribute 12TH_MARKS INTEGER
@attribute CGPA REAL
@attribute TNOA INTEGER
@attribute YEARGAP INTEGER
@attribute RESULT {A,B,C,F}

@data
M,21,CBSE,CBSE,91,90,8.51,0,0,A
M,22,SB,SB,75,91,8.43,0,0,A
F,21,CBSE,CBSE,91,67,7.68,1,0,B
M,23,SB,SB,80,74,7.4,1,1,B
M,21,ICSE,CBSE,87,82,8.51,0,0,A
M,21,CBSE,SB,74,91,8.32,0,0,B
M,23,ICSE,CBSE,89,92,7.6,2,1,C
M,22,SB,CBSE,80,67,8.4,1,0,B
F,21,CBSE,CBSE,90,90,9,0,0,A
F,21,ICSE,ICSE,90,90,8,1,1,B
M,22,CBSE,CBSE,60,70,8,0,1,B
M,21,CBSE,CBSE,70,70,6,4,0,F
M,21,CBSE,ICSE,80,80,7,1,0,B
M,22,CBSE,ICSE,80,80,7.52,0,1,B
M,21,ICSE,ICSE,80,85,8.52,0,0,A
M,21,ICSE,ICSE,75,80,6.01,2,0,C
M,21,CBSE,SB,72,67,6.56,3,0,C
F,20,CBSE,CBSE,72,76,7.96,2,0,B
F,20,ICSE,SB,80,87,8.34,1,0,B
```

Fig. 1. Dataset

Steps to be performed for implementing the two algorithms in WEKA too, are as follows:

Step-1: First of all, import the dataset into the tool and preprocess it. Screen below shows the preprocessing step:

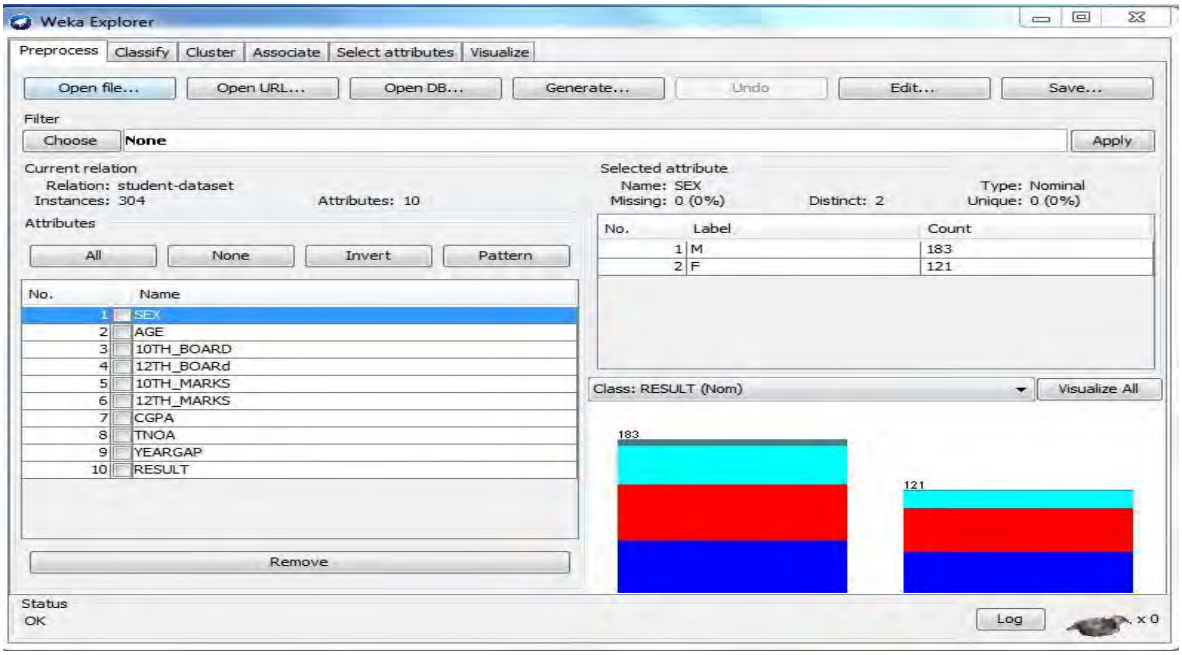


Fig. 2. Preprocessing

Left hand side of the above screen shows detail of relation name, number of attributes and number of records. Right hand side gives details of attribute values , type, missing values and number of distinct values. Specification of every attribute is displayed in the right bottom of the screen.

Step-2: Classify the data using Naïve Bayers classification algorithm. Following screen shows result obtained from Naïve Bayers classification algorithm. Use percentage split to divide dataset into two(70% training data,30%test data)

```
Time taken to build model: 0.02seconds

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

Correctly Classified Instances         228               75      %
Incorrectly Classified Instances        76               25      %
Kappa statistic                          0.6211
Mean absolute error                      0.1724
Root mean squared error                  0.3167
Relative absolute error                 51.5238 %
Root relative squared error             77.5001 %
Total Number of Instances              304

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.847     0.062      0.887     0.847      0.866      0.928    A
                 0.805     0.226      0.693     0.805      0.745      0.833    B
                 0.5       0.08       0.635     0.5        0.559      0.84     C
                 0.667     0.01       0.667     0.667      0.667      0.986    F
Weighted Avg.    0.75      0.128      0.75      0.75       0.747      0.874

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 94 12  5  0 |  a = A
 12 95 11  0 |  b = B
  0 30 33  3 |  c = C
  0  0  3  6 |  d = F
```

Fig. 3. Results of Naive Bayes Algorithm

Step-3: Delete previous buffer value and predict using J48 decision tree algorithm.

Step-4: Classify the data using J48 decision tree algorithm. Following screen shows result obtained from J48 decision tree algorithm. Use same percentage split to divide dataset into two(70% training data,30%test data). Following screen shows results obtained after applying J48 algorithm:

```
                CGPA
                TNOA
                YEARGAP
                RESULT
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48graft pruned tree
------------------

CGPA <= 8.41
|   TNOA <= 2
|   |   CGPA <= 7.94: C (60.0/20.0)
|   |   CGPA > 7.94: B (109.0/19.0)
|   TNOA > 2
|   |   CGPA <= 6.53
|   |   |   10TH_BOARD = ICSE: F (3.0)
|   |   |   10TH_BOARD = CBSE: C (4.0/1.0)
|   |   |   10TH_BOARD = SB: F (3.0)
|   |   CGPA > 6.53: C (7.0)
CGPA > 8.41: A (118.0/13.0)

Number of Leaves  :     7

Size of the tree :      12
```

Fig. 4.Tree generated by J48 Algorithm

The following snapshot specifies the overall datset.

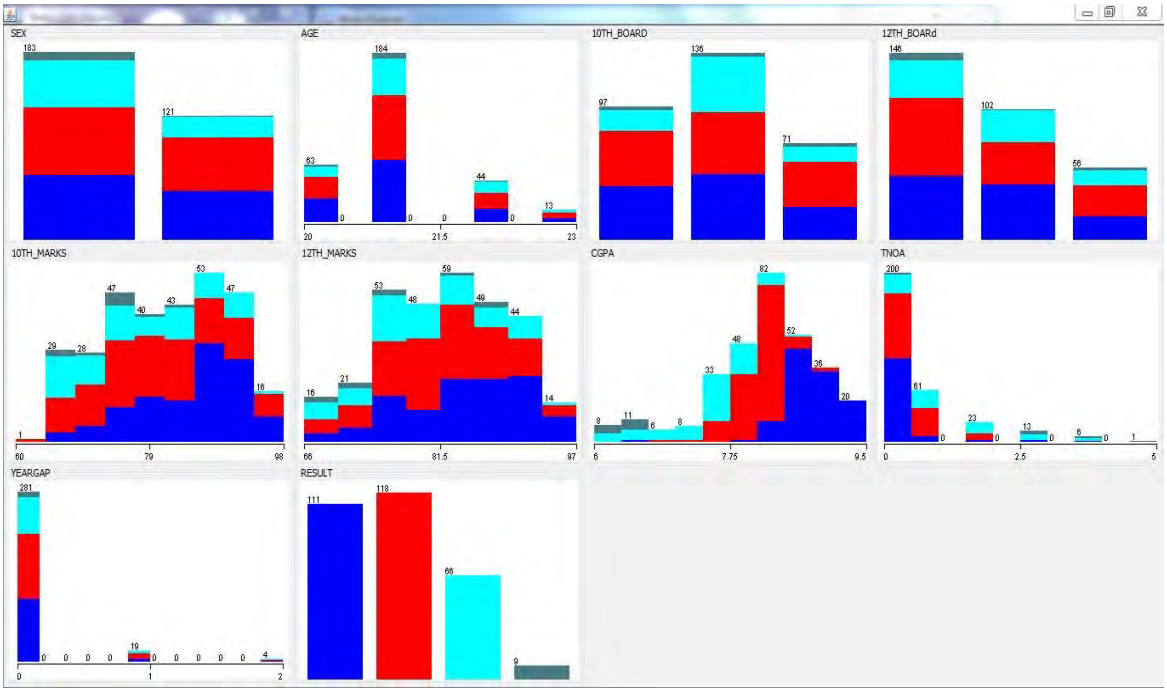


Fig. 5.Visualisation

Implementation of wID3 algorithm is done in C programming language. The output of the program shows the rules generated by the algorithm for both the classes, that is, for class PASS and class FAIL

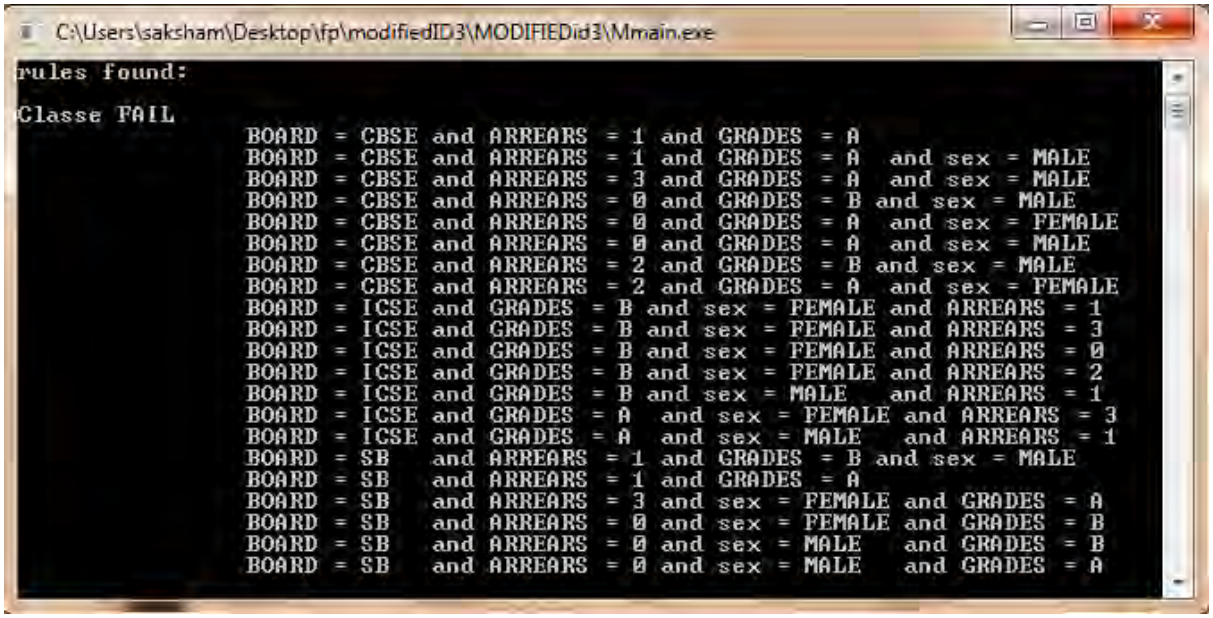


Fig. 6.Rules for class 'FAIL'

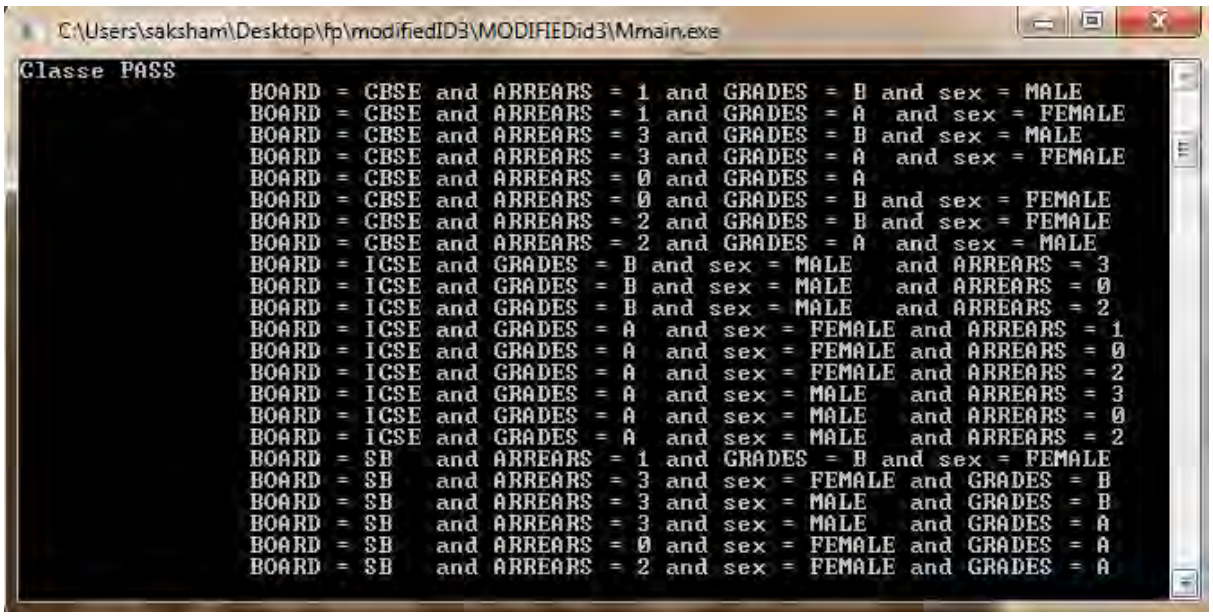


Fig. 7.Rules for class 'PASS'

VIII. Conclusion and Future Enhancement

The improved weighted ID3 algorithm based on Gain Ratio decided whether the attributes should be modified or not by checking its weight. So, the inclination problem presented in traditional ID3 problem is overcome. wID3's performance is also compared with J48 algorithm and Naïve Bayes classification algorithm. By doing performance analysis of all the three algorithms, we came to a result that wID3 algorithm is more efficient than other two algorithms. wID3 classified 93% of records accurately, whereas J48 and Naïve Bayes classified only 78.6% and 75% records respectively. In future, we can build a user-friendly softwares of teachers using the concept of wID3 algorithm in which a teacher will just have to enter the details of the students and it will display the possible result of those students in term-end exams. Also a flexible system can be made in which we can add the dataset dynamically and the system updates its results automatically.

References

[1] B.K. Bharadwaj and S. Pal. "Mining Educational Data to Analyze Students Performance", International Journal of AdvanceComputer Science and Applications (IJACSA), Vol. 2, No. 6, pp.63-69, 2011.

[2] P.K.Srimani and Annapurna S Kamath."Data Mining Techniques for the Performance Analysis of a Learning Model-A case study"International Journal of Computer Applications(0975-8887),volume 53-No 5 September 2012.

[3] Surjeet Kumar Yadav,Bhardwaj and S Pal."Data Mining Applications: A Comparative Study for Predicting Student's performance"International Journal of Innovative Technology & Creative Engineering(ISSN;2045-711)Vol.1 no.12 December.

[4] C.MARQUEZ-VERA , C.ROMERO and S.VENTURA "Predicting School Failure Using Data Mining"2011

[5] Sajadin Sembiring ,M.Zarlis ET.AL ."Prediction of Student Academic Performance By an Application of Data Mining Techniques"2011 International Conference on Management and Artificial Intelligence IPEDR vol.6(2011)IACSIT press,Bali,Indonesia
[6] K.Nandhiini and S.Saranya "ID3 Classifier for pupils Status Prediction"International Journal of Computer Application(0975-8887)vol 57-No.3,November 2012.
[7] R.R.Kabra and R.S.Bichkar."Performance of Engineering Students using Decision Trees".International Journal of Computer Applications(0975-8887)volume 36-No.11,December 2011
[8] Mrinal Pandey and Vivek Kumar Sharma."A Decision Tree Algorithm Pertaining to the Student Performance Analaysis and Prediction"".International Journal of Computer Applications(0975-8887)volume 61-No.13,January 2013
[9] Dorina Kabakchieva "Student Performance Prediction by using Data Mining Classification Algorithms"".International Journal of Computer Science and Management Research Vol issue 4 November 2012 ISSN 2278-733X
[10] Zlatkok J.Kovacic "predicting Success by Mining Enrolment data".Research in Higher Education Journal.