Calculating Students Performance by Implementing ID3 Algorithm

1st Md. Mahidul Hasan
Dept. of CSE
Varendra University
Rajshahi, Bangladesh
mahidulhasanbd@gmail.com

2nd Md. Meraj Ali
Dept. of CSE,
Varendra University
Rajshahi, Bangladesh
meraj09034@gmail.com

3rd Hadisur Rahman
Dept. of CSE,
Varendra University
Rajshahi, Bangladesh
hudacse6@gmail.com

***Abstract*— Now a days it's become a big challenge to determine student's quality of an educational institution. An educational institution needs to have an approximate prior knowledge of enrolled students to predict their performance in future academics. This helps them to identify promising students and also provides them an opportunity to pay attention to and improve those who would probably get lower grades. As a solution, we have worked which can predict the performance of students from their previous performances using concepts of data mining techniques under ID3 algorithm. We have analyzed the data set containing information about students, such as their previous Semester results, Study hours, Internet browsing hours, Playing hours, Watching Television hours etc. And finally we have reached a decision that what performance will acquire by a students. From this research we will be able to find out students' future movement. This research will help students to take proper steps to reach their goal smoothly.**

Keywords—Data mining, ID3 algorithm, Classification, Decision tree.

I. INTRODUCTION

Data mining is the tools of collecting, searching through, and analyzing a large amount of data in a database, as to discover patterns or relationships. Data mining are very important in the classification of the objects. That is why it has been successfully used in expert systems in capturing knowledge. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. In this paper the ID3 decision tree learning algorithm is implemented with the help of an example which includes 50 individuals among 17 which were very good students, 26 were good students and 07 were bad students. The basic calculations are used to calculate the classification related to the training set used. The resultant of the work will be the classified decision tree and the decision rules. This paper focuses on comparing the performance accuracy of ID3 techniques of the decision tree for calculating students' performance using WEKA Based on the characteristics of this algorithm, the main objectives of the research work is to construct the decision tree until the appropriate classification is reached.

II. METHODOLOGY AND ALGORITHM

A. *How Much Data We Used?*

- Here I have used 50 individual data set that I collected from the students of Varendra University.
- I have collected data XAMPP through web pages.
- Here I have collected more than 100 set of data and used 50 set of data as calculation.

B. *Working on Algorithm's*

ID3- The main method used in this paper is the ID3 algorithm, which is used for the classification and the pattern reorganization technique. The paper is based on many of the survey of the journals and the publications in the field of the data mining. Data mining techniques basically use the ID3 algorithm as it's the basic algorithm of classification. The first work on ID3 was done by J.R Quinlan in 1986. [2]

- Data Classification by Naive Bayes.
- Data Classification by J48
- Data Classification by Random Tree.

C. Using Software

- Weka
- Xampp
- MS Excel

D. *Pc Configuration where the data is tested.*

a) 08GB RAM
b) 120GB SSD
c) Core I5(3.6GHZ) processor
d) Window 10 pro (64 bit)

E. *Learning Set:*

First categorize the value of the above table for better calculation result using ID3 algorithm. Table below indicates the acceptable value.

Table 01: Categorized attributes and Values

Attribute	Possible Values		
Study Hour	Very little(SVL) (t<2)h	Medium(SM) (2<=t<4)h	Very High(SVH) (t>=4)h
Playing Hour	Little(PL) (t<1)h	Good(PG) (1<=t<2)h	High(PH) (t=>2)h
Internet Browsing Hour	Little(IL) (t<3)h	Good(IG) (3=<t<6)h	High(IH) (t>=6)h
Watching TV Hour	Low(WL) (t<1)h	Medium(WM) (1<=t<2)h	High(WH) (t=>2)h
SSC & HSC & B.Sc (AVG)	RC+ (M<4.00)	RB+ (4.00<=M<4.50)	RA+ (M>=4.50)
Summary	Bad(B)	Good(G)	Very Good (VG)

F. Data Calculation*:* **Step by Step Calculation**

STEP (1-3.3): Root Calculation

First calculate the Entropy(S) of the example set. Our example set contains 17 Very Good, 26 Good, 7 Bad.

Entropy(S) $= -\frac{17}{50}\ log_2(^{17}/_{50}) - \frac{26}{50}\ log_2(^{26}/_{50}) - \frac{7}{50}\ log_2(^{7}/_{50})$ =0.529174+.490577+.39711

= 1.416861

STEP 2.1: Study Hour (SHr)

Study Hour value can be Very little (SVL), Medium (SM), High (SH). Total frequency of Very little (SVL), Medium (SM), High (SH), are 22, 20 and 8 respectively. Where,
Study Hour = Very little (SVL), 8 of subjects are Very Good, 11 of subjects are Good, 3 of subjects are Bad.
Study Hour = Medium (SM), 8 of subjects are Very Good, 9 of subjects are Good, 3 of subjects are Bad.
Study Hour = High (SH), 1 of subjects are Very Good, 6 of subjects are Good, 1 of subjects are Bad.

Entropy (Study Hour, SVL) $= -\frac{8}{22}\ log_2(^{8}/_{22}) - \frac{11}{22}\ log_2(^{11}/_{22}) - \frac{3}{22}\ log_2(^{3}/_{22})$ =1.422674

Entropy (Study Hour, SM) $= -\frac{8}{20}\ log_2(^{8}/_{20}) - \frac{9}{20}\ log_2(^{9}/_{20}) - \frac{3}{20}\ log_2(^{3}/_{20})$

=1.457716

Entropy (Study Hour, SH) $= -\frac{1}{8}\ log_2(^{1}/_{8}) - \frac{6}{8}\ log_2(^{6}/_{8}) - \frac{1}{8}\ log_2(^{1}/_{8})$

=1.061278

Gain **(S,** Study Hour**)** = Entropy(S) – (22/50) x Entropy (Study Hour SVL)– (20/50) x Entropy (Study Hour, SM)– (8/50) x Entropy (Study Hour, SH)
= 1.416861– (22/50) x 1.422674 – (20/50) x 1.457716 – (8/50) x 1.061278

= 0.03799

STEP 2.2: Attribute Playing Hour (PHr)

Playing Hour can be Little, Good and High. Total frequency of Little (PL), Good (PG), and High (PH) are 11, 25 and 14 respectively. Where,

Playing Hour = Little (PL), 2 of subjects are Bad, 7 of subjects are Good, 2 of subjects are Very Good.

Playing Hour = Good (PG), 4 of subjects are Bad, 11 of subjects are Good, 10 of subjects are Very Good.

Playing Hour = High (PH), 1 of subjects are Bad, 8 of subjects are Good, 5 of subjects are Very Good.

Entropy (Playing Hour, PL) $= -\frac{2}{11}\ log_2(^{2}/_{11}) - \frac{7}{11}\ log_2(^{7}/_{11}) - \frac{2}{11}\ log_2(^{2}/_{11})$ =1.309295

Entropy (Playing Hour, PG) $= -\frac{10}{25}\ log_2(^{10}/_{25}) - \frac{11}{25}\ log_2(^{11}/_{25}) - \frac{4}{25}\ log_2(^{4}/_{25})$ =1.472935

Entropy (Playing Hour, PH) $= -\frac{5}{14}\ log_2(^{5}/_{14}) - \frac{8}{14}\ log_2(^{8}/_{14}) - \frac{1}{14}\ log_2(^{1}/_{14})$

=1.263810

Gain **(S,** Playing Hour**)** = Entropy(S) – (11/50) x Entropy (Playing Hour, PL)– (25/50) x Entropy (Playing Hour, PG) – (14/50) x Entropy (Playing Hour, PH)
= 1.416861– (11/50) x 1.309295– (25/50) x 1.472935– (14/50) x 1.263810

= 0.038481

STEP 2.3: Attribute Internet Browsing Hour (IB)

Internet Browsing Hour can be Little, Good and High. Total frequency of Little (IL), Good (IG), and High (IH) are 4, 18 and 28 respectively. Where,

Internet Browsing Hour = Little (IL), 1 of subjects are Bad, 3 of subjects are Good, 0 of subjects are Very Good.

Internet Browsing Hour = Good (IG), 1 of subjects are Bad, 9 of subjects are Good, 8 of subjects are Very Good.

Internet Browsing Hour = High (IH), 5 of subjects are Bad, 14 of subjects are Good, 9 of subjects are Very Good.

Entropy (Internet Browsing, IL) $= -\frac{1}{4}\ log_2(^{1}/_{4}) - \frac{3}{4}\ log_2(^{3}/_{4}) - \frac{0}{4}\ log_2(^{0}/_{4})$ =0.811278

Entropy (Internet Browsing, IG) $= -\frac{1}{18}\ log_2(^{1}/_{18}) - \frac{9}{18}\ log_2(^{9}/_{18}) - \frac{8}{18}\ log_2(^{8}/_{18})$ =1.251626

Entropy (Internet Browsing, IH) $= -\frac{5}{28}\ log_2(^{5}/_{28}) - \frac{14}{28}\ log_2(^{14}/_{28}) - \frac{9}{28}\ log_2(^{9}/_{28})$ =1.470141

Gain **(S,** Internet Browsing**)** = Entropy(S) – (4/50) x Entropy (Internet Browsing, IL)– (18/50) x Entropy (Internet Browsing, IG) – (28/50) x Entropy (Internet Browsing, IH)
= 1.416861– (4/50) x0.811278– (18/50) x 1.251626 – (28/50) x 1.470141

= 0.078095

STEP 2.4: Attribute Watching TV Hour (WT)

Watching TV Hour can be Low, Medium and High. Total frequency of Low (WL), Medium (WM), and High (WH) are 21, 18 and 11 respectively. Where,

Watching TV Hour = Low (WL), 3 of subjects are Bad, 11 of subjects are Good, 7 of subjects are Very Good.

Watching TV Hour = Medium (WM), 4 of subjects are Bad, 8 of subjects are Good, 6 of subjects are Very Good.

Watching TV Hour = High (WH), 0 of subjects are Bad, 7 of subjects are Good, 4 of subjects are Very Good.

Entropy (Watching TV, WL) $= -\frac{3}{21}\ log_2(^{3}/_{21}) - \frac{11}{21}\ log_2(^{11}/_{21}) - \frac{7}{21}\ log_2(^{7}/_{21})$ =1.418024

Entropy (Watching TV, WM) $= -\frac{4}{18}\ log_2(^{4}/_{18}) - \frac{8}{18}\ log_2(^{8}/_{18}) - \frac{6}{18}\ log_2(^{6}/_{18})$ =1.530491

Entropy (Watching TV, WH) = - $\frac{0}{11}$ $log_2(^0/_{11})$ - $\frac{7}{11}$ $log_2(^7/_{11})$ - $\frac{4}{11}$ $log_2(^4/_{11})$

=0.945658

Gain **(S,** Watching TV**)** = Entropy(S) – (21/50) x Entropy (Watching TV, WL)– (18/50) x Entropy (Watching TV, WM) – (11/50) x Entropy (Watching TV, WH)

= 1.416861– (21/50) x1.418024– (18/50) x 1.530491– (11/50) x 0.945658

= 0.062269

STEP 3: Summary of results are listed below.

Entropy(S)A	= 1.416861
Gain(S**,** Study Hour)	= 0.037994
Gain(S, Playing Hour)	= 0.038481
Gain(S, Internet Browsing)	= 0.078095
Gain(S, Watching TV)	= 0.062269

To find which attribute is the root node, we just have taken the highest Gain value.

Gain (S, Internet Browsing Hour**)** = 0.078095 is the highest among all attribute. Therefore**, "Internet Browsing"** is the decision attribute in the root node. So, after calculating this step the decision tree contains only the root node that is Internet Browsing. We can depict the decision tree as follows.

Fig. 01. Root Node

III. RESULTS

STEP 4: Summary of results are listed below.

WL	
Entropy(S)	= 1.418024
Gain(S, WL,SHr)	= 0.049002
Gain(S, WL,PHr)	= 0.036169

WM	
Entropy(S)	= 1.530491
Gain(S, WM,SHr)	= 0.174582
Gain(S, WM,PHr)	= 0.306098

WH	
Entropy(S)	= 0.945658
Gain(S, WH,SHr)	= 0.355638
Gain(S, WH,PHr)	= 0.072054

Fig. 2.

STEP 5: Summary of results are calculated by weka is given below.

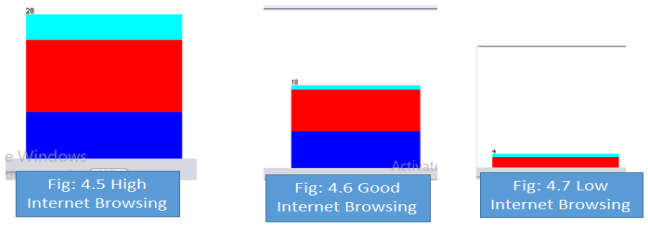

Fig. 3.

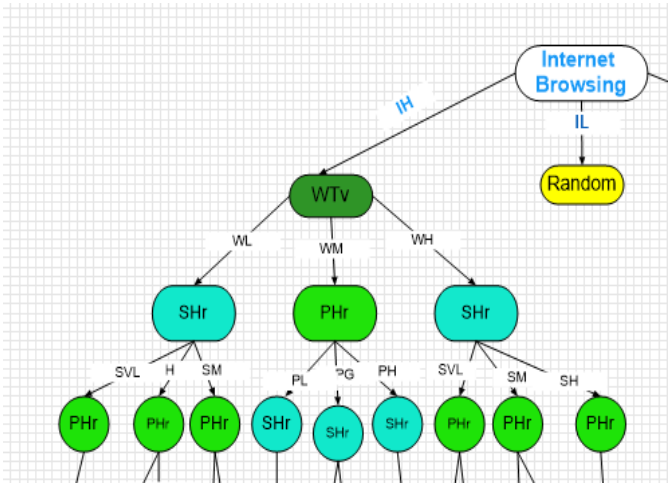

Fig. 4.

STEP 6: Summary of results are listed below.\

PG	
Entropy(S)	= 1.472935
Gain(S, PG,SHr)	= 0.039638
Gain(S, PG,WTv)	= 0.061569

PH	
Entropy(S)	= 1.263810
Gain(S, PH,SHr)	= 0.432456
Gain(S, PH,WTv)	= 0.333843

PL	
Entropy(S)	= 1.302995
Gain(S, PL,SHr)	= 0.218387
Gain(S, PL,WTv)	= 0.473958

Fig. 5.

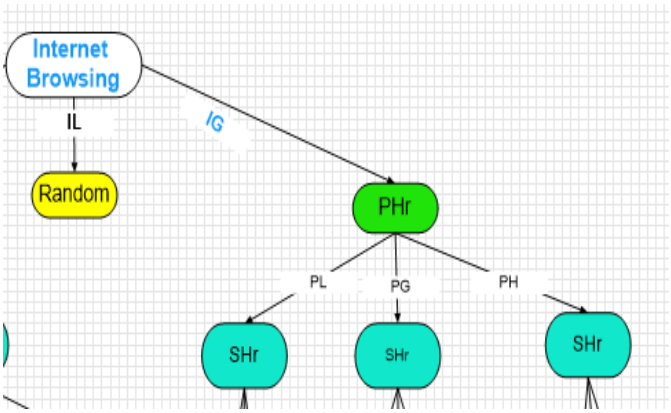

Fig 06

STEP 7: Summary of results are calculated by weka is given below

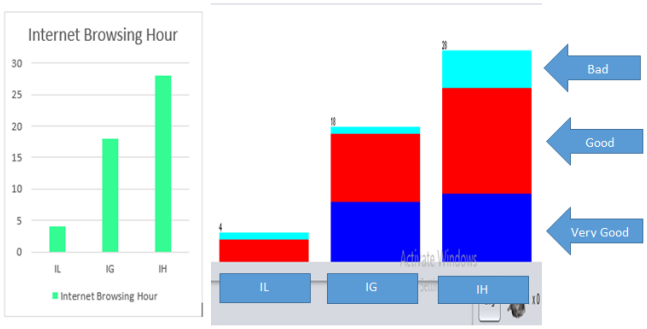

Fig. 7.

Here from the above figures (1-7) we can observe that, here in the current relation number of 50 instances is affected and also number of 5 attributes are affected.

we can see from attribute attributes section we have five attributes, For the first attribute named "Internet Browsing Hour' we observe that it has three distinct branch names IL, IG and IH.

Now from summary view part we can see that for IL or Less Internet Browsing it is affected only number of 4 data correctly, IG or Good Internet Browsing which is affected number of 18 data correctly and IH or High Internet Browsing which is affected number of 28 data correctly.

STEP 8: Summary of results are calculated by weka is given below

After fully calculation we have this final decision tree.

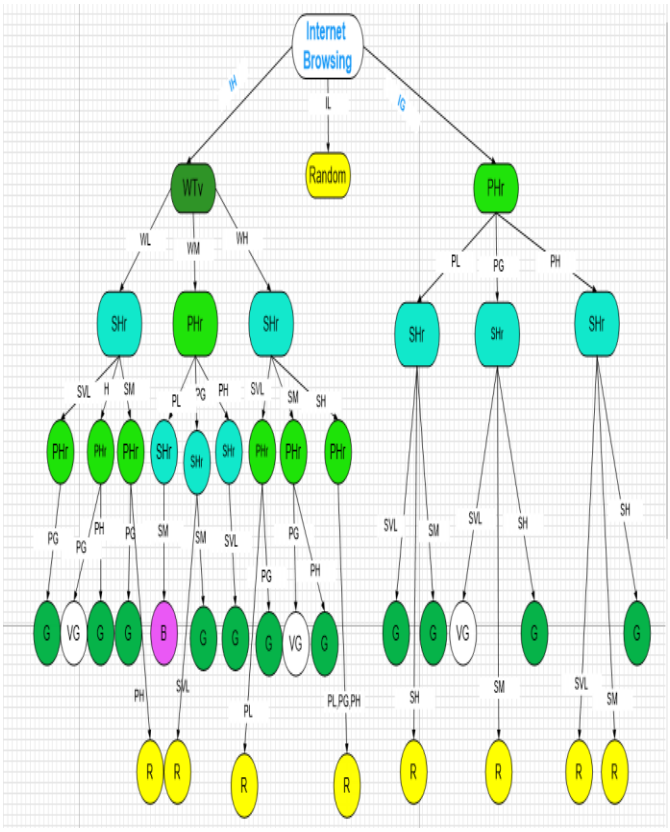

Fig. 8. Final Decision tree after step step calculation by

IV. CONCLUSION

In this research, the ID3 algorithm is used to evaluate student's performance and as there are many approaches that is used for data classification and the decision tree method is used here. Student's information like Study Hour, Internet Browsing Hour, watching television hour, Playing hour and students previous results were collected from the student's, to predict the future performance of the students. Data Mining is gaining its popularity in almost all applications of real world. One of the data mining techniques that is, ID3 is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce clear and soft rules that are easy to prediction than other methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best methods for Student data to predict the student's performance in the future.

V. FUTURE WORK

To get more perfect visualize tree we need to classify data more efficiently and also analyses the data vary carefully. Again, we need to calculate full process to get more perfect results for calculating student's future performance.

VI. REFERENCES

[1] Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques, Elsevie

[2] Dunham, M.H., (2003) *Data Mining:* Introductory and Advanced Topics, Pearson Education Inc.

[3] Sunitha soni, Associate professor "Predictive Data mining For Medical Diagnosis an Overview of Heart Dis Ease March 2011". [4] D.Senthil Kumar, "Decision Support System for Medical Diagnosis Using Data Mining March 2011".

[4] Devroye L, Gyorfi L, Lugosi G (1996) A probabilistic theory of patternrecognition. Springer, New York. ISBN 0-387-94618-7J

[5] Vıctor.H.Garcıa, Raul Monroy and Maricela Quintana, "Web Attack DetectionUsingID3".Ahmad Baraani-Dastjerdi; Josef Pieprzyk;

[6] Sonika Tiwari and Prof. Roopali Soni, "Horizontal partitioning ID3 algorithm A new approach of detecting network anomalies using decision tree", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 7, September – 2012..

[7] Thuraisingham, B. Big data security and privacy. In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, San Antonio, TX, USA, 2–4 March 2015; pp. 279–280.

[8] Kumar Ashok, Taneja H C, Chitkara Ashok K and Kumar Vikas, "Classification of Census Using Information Theoretic Measure Based ID3 Algorithm

[9] C. Apte and S. Weiss, "Data Mining with Decision Trees and Decision Rules", Future Generation Computer Systems, vol. 13, (1997), pp. 197-210

[10] Tina R. Patil, Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," International Journal of Computer Science and Applications, Volume 6, No.2, Apr 2013

[11] Yusuf, U. & Gülay, T., Rule learning with Machine Learning Algorithms and Artificial Neural Network Journal of Selcuk University Naurral Applied Scince, vol.1,no.2, 2012.

[12] G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to Knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in2001.