iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Student Prediction System Using Data Mining Algorithm- Survey

Ravi Kumar Rathore[1], J. Jayanthi[2], Gurpreet Kaur[3]
[1,2,3] *Lovely Professional University Phagwara, Punjab, India*

Abstract: ***Student prediction system proposal is efficient approach for classifying student data based on merit. Classifying students based on merit is a tedious task when the student number and concern subject are high for placements. We propose student data classification using many approaches like Decision Tree, C4.5 algorithms, ID3 algorithm, Genetic Algorithm, and Neural Network. This may be lead to efficient use of student education data base for placement and non placement classes.***

Keywords: Data Mining, Classification, Decision Tree, ID3, C4.5, Genetic Algorithm, Neural Network, Prediction, Student Data set.

I. INTRODUCTION

Classification is most frequent technique which is used for classify the data set, data set it may be any kind of data set such as student data in educational field, public data for classify them for something, and many other kind of data. Classification technique can apply on any kind of data set for predicting something. Basically classification technique is map the input data into output data or result according to your requirement. Now a day's classification is using in many different-different areas like education, industrial, medical and other many places [1]. Classification is basically data mining technique in which apply some input pattern and get desire output by using any one classification algorithm. Classification is supervised learning which require creating or generating some rules for classifying test data into the predefined classes. This classification technique requires some phases such as the first phase is learning process and second one is to analyze given data ser or given input data and classification rules generated.

II. LITERATURE SURVEY

In the survey of literature, we have seen that many researchers done research for calculating performance of the student, judge intelligence of the student and predict them for something [2, 3].

There exists numerous algorithm produced to construct classification model for student prediction [14] such as J48 decision tree algorithm to predict students' final GPA based on their grades in previous courses and collected transcripts data of female students who graduated from Computer Sciences in King Saud University in the year 2012[Mashael A. Al-Barrak and Muna Al-Razgan], [15] C4.5 and ID3 to predict the student performance and collected first year student data of Fr. C.R.I.T., Navi Mumbai [Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul], [16] Decision tree algorithm for to analyze large amounts of risk factors that can affect student's performance in exams and collected by different surveys from Shivaji University, Kolhapur[Priyanka A. Patil,R. V. Mane]. [17] Decision tree: C4.5 algorithm, ID3 algorithm CART algorithm to predict student performance in the final exam and data set collected by them from VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method for Institute of Engineering and Technology for session 2010[Surjeet Kumar Yadav, Saurabh Pal]. Decision Tree to predict the student's performance in higher education. [Krina Parmar, Prof. Dineshkumar Vaghela and Dr Priyanka Sharma].

III. CLAASIFICATION METHODOLOGIES

A. Desion Tree

Basically Decision Tree is the Data Mining technique which used for classification[4,5]. It is like tree which have root, internal node and leaf node also. In the structure of decision tree internal node shown by rectangle and each leaf node in the decision tree represented by ovals. When we test the data set in that case internal node splits according to maximum output, internal node is used for testing the parameters of given attribute. This tree approach is used for gaining the information and splitting internal node on the basis of some important calculations. In this technique leaf node show the final outcome. There are two most common techniques which is used for creating decision tree: ID3 and C4.5[6,7].

B. Id3

ID3[8,10] is the one of the most important method or technique which is used for making decision tree. ID3 is basically a classification technique which use for making or generating a decision tree. ID3 is an algorithm which use for creating Decision Tree, ID3 calculate information gain and measure the to choose the splitting attribute to making decision tree so it is very important approach for making decision tree. ID3 Algorithm chooses the selected attributes for building decision tree structure. It cannot give exact result when there is noise and missing attribute because this algorithm does not handle the missing attribute. For handling and removing these terms we use C4.5. ID3 is not able to handle missing attribute.

Some important calculations are there to making a decision tree, we use some formulae to making a decision tree and splitting criteria.

C. C4.5

C4.5 algorithm is an extension of ID3 algorithm [15] which developed by Quinlan Ross. C4.5 handles both noisy and missing attributes to making a decision tree. To handle missing and noisy attributes we use C4.5 algorithm. C4.5 splits the attribute values into two partitions based on the some calculations of entropy and information gain such that all the values of the tree represents child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to making a decision tree. C4.5 algorithm removes the biasness of information gain when there are many result values of an attribute. First, calculate the gain ratio of each attribute. The root node will be the attribute whose information gain ratio is maximum and this process again and again happens until unless final outcome does not come. C4.5 uses pruning method to remove unnecessary branches in the decision tree to improve the accuracy and for get performance of classification.

D. Neural Network

Neural Networks concept is proposed on the basis of working human brain. In artificial neural network there are three thinks, first one is input vector, weight and output vector. In artificial neural network when we trained our data then there is a concept learning rate parameter, it denoted by eta its value belongs between 0 and 1. If value of learning rate is 1 which is high in that case then our system is more learnable and it improve the performance of the function if its value is 0 then it is less learnable and performance of the function will low. When we compute the output then there is a factor called error we can minimize the error by adjusting the weight. We can adjust the weights at output layer and also on hidden layer. We can use back propagation algorithm for multilayer perceptron. For minimizing the error we use some strategy for classification in neural network such as: BPA and any other[11].

E. Genetic Algorithm

This approach is most important method. Genetic algorithm uses two some method for solving some problems in genetic algorithms, problems are constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. In this algorithm select population for solving problem. Genetic algorithm is a heuristic search in the artificial intelligence which mimics the process of natural selection. In the genetic algorithm the heuristic is also sometimes called a Meta heuristic strategy is used to make solutions to optimization and search problems. Genetic algorithms have larger class of evolutionary algorithms. Generate solutions is used to optimization problems using some techniques or methods inspired by natural evolution, which is defined under as follows: mutation, crossover and fitness function. By using this approach in classification we can get accurate result and we can also minimize the error.

IV. EXAMPLE OF ID3 ALGORITHM

Some important calculation are there to making a decision tree, we use some formulae to making a decision tree and splitting criteria.:Entropy and Information Gain.

Here taking a example of student data set for making decision tree using ID3 algorithms.

Entropy = - Σi Pi log2Pi

Entropy is the calculation of positive (+) and negative examples (-).

A. Information Gain

Information Gain use for calculating best fit attribute for every node and best attribute become the root node. Suppose the information has Gain (G, A) of an attribute A, it is to a collection of examples in G, which is defined under as follows:

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Gain(G,A) = Entropy (G) - ∑ |Gv|÷|G| Entropy (Gv)

∑ = Value (A) is the set of all possible values for attribute A.

Gv = Gv is the subset of G for which Attribute A has v , i.e., Gv = {g = Gv | A(g) = v }).

The first calculation or term in information Gain is the entropy of the original collection *G* and the second term is the desired value of the entropy after G is partitioned using attribute *A*. The expected entropy described by second term is the sum of the entropies of each subset, the fraction of examples |Gv|÷|G| that belong to Gain (G, A) is therefore the optimal reduction in entropy happen by knowing the value of attribute A.

B. Splitting Criteria

Split Information (G,A) = - ∑ |Gi|÷|G| log2 |Gi|÷|G|

Where ∑ is(i=1 to n) And

Gain Ratio (G,A) = Gain (G,A) ÷ Split Information (G,A)

C. Some terminology and rules for classifying attributes are

If the entropy of the attribute is 0, it relates to same types node and there is no need to classify further more. If the entropy of the attribute is 1, it relates to different types node and there is a need to classify further more.

Table 1: student data set.

No.	Student name	10th Marks(%)	12th Marks(%)	B.Tech C.G.P.A.	M.Tech C.G.P.A.	Placement Non Place-ment training
1	Nupur Ahluwalia	High	High	High	High	Yse
2	Ravi Kumar	Med	Low	Med	Med	Yse
3	Vanshika Shrivas-tav	Med	Med	Med	Med	Yse
4	Gaurav Dixit	High	Med	Med	Med	Yse
5	Sachin Singh	Low	Med	Med	High	No
6	Ravindra singh	Med	Low	Med	Med	Yes
7	Neha	Low	Med	Med	Med	No
8	Seenam	Low	Med	High	High	Yes
9	Kunal Gupta	Med	Med	High	High	Yes
10	Ravi Rathore	Low	Med	Med	High	Yes
11	Swetank	Low	Low	Low	Med	No
12	Narendra Yadav	Low	Med	Med	Low	No

1) *Step 1 :* Example set S

Let we have a set S of 12 examples with 8 “Yes” and 4 “No” then

Entropy (S) = - (8/12) log2 (8/12) – (4/12) log2 (4/12) = 0.91829

2) *step 2 :* Attribute B.Tech

In B.Tech attribute there are three possible values such as Low, Medium and High.

B.tech = Low is of occurrence 2

B.tech = Medium is of occurrence 7

B.tech = High is of occurrence 3

B.tech = In this Low, out of 2 there are 0 'Yes' and 2 'No'

B.tech = In this Medium, out of 7 there are 4 'Yes' and 3 'No'

B.Tech = In this High, out of 3 there are 3 'Yes' and 0 'No'

Entropy (Low) = - (0/2) log2 (0/2) – (2/2) log2 (2/2) = 0

Entropy (Medium) = - (4/7) log2 (4/7) – (3/7) log2 (3/7) = 0.98522

Entropy (High) = - (3/3) log2 (3/3) – (0/3) log2 (0/3) = 0

Gain (S, B.Tech) = Entropy (S) – (2/12) × Entropy (Low) – (7/12) × Entropy (Medium) – (3/12) × Entropy (High)

= 0.91829– (2/12) × 0 – (7/12) × 0.98522 – (3/12) × 0

= 0.91829 – 0.57471

= 0.34358

3) *STEP 3 :* Attribute M.Tech

In M.Tech attribute thare are three possible values like Low, Medium and High.

M.tech = Low is of occurrence 0

M.tech = Medium is of occurrence 7

M.tech = High is of occurrence 5

M.tech = In this Low, out of 0 there are 0 'Yes' and 0 'No'

M.tech = In this Medium, out of 7 there are 4 'Yes' and 3 'No'

M.tech = In this High, out of 5 there are 4 'Yes' and 1 'No'

Entropy (Low) = - (0/0) log2 (0/0) – (0/0) log2 (0/0) = 0

Entropy (Medium) = - (4/7) log2 (4/7) – (3/7) log2 (3/7) = 0.98522

Entropy (High) = - (4/5) log2 (4/5) – (1/5) log2 (1/5) = 0.72192

Gain (S, M.Tech) = Entropy (S) – (0/12) × Entropy (Low) – (7/12) × Entropy (Medium) – (5/12) × Entropy (High)

=0.91829– 0 – 0.574711 – 0.3008

= 0.042779

B.Tech attribute has highet gain so that, it will become the decision node.

4) *Step 4 :* This process repeatedly going on until all data do not classify perfectly or we run out of attributes.

5) *step 5 :* Making Decision Tree

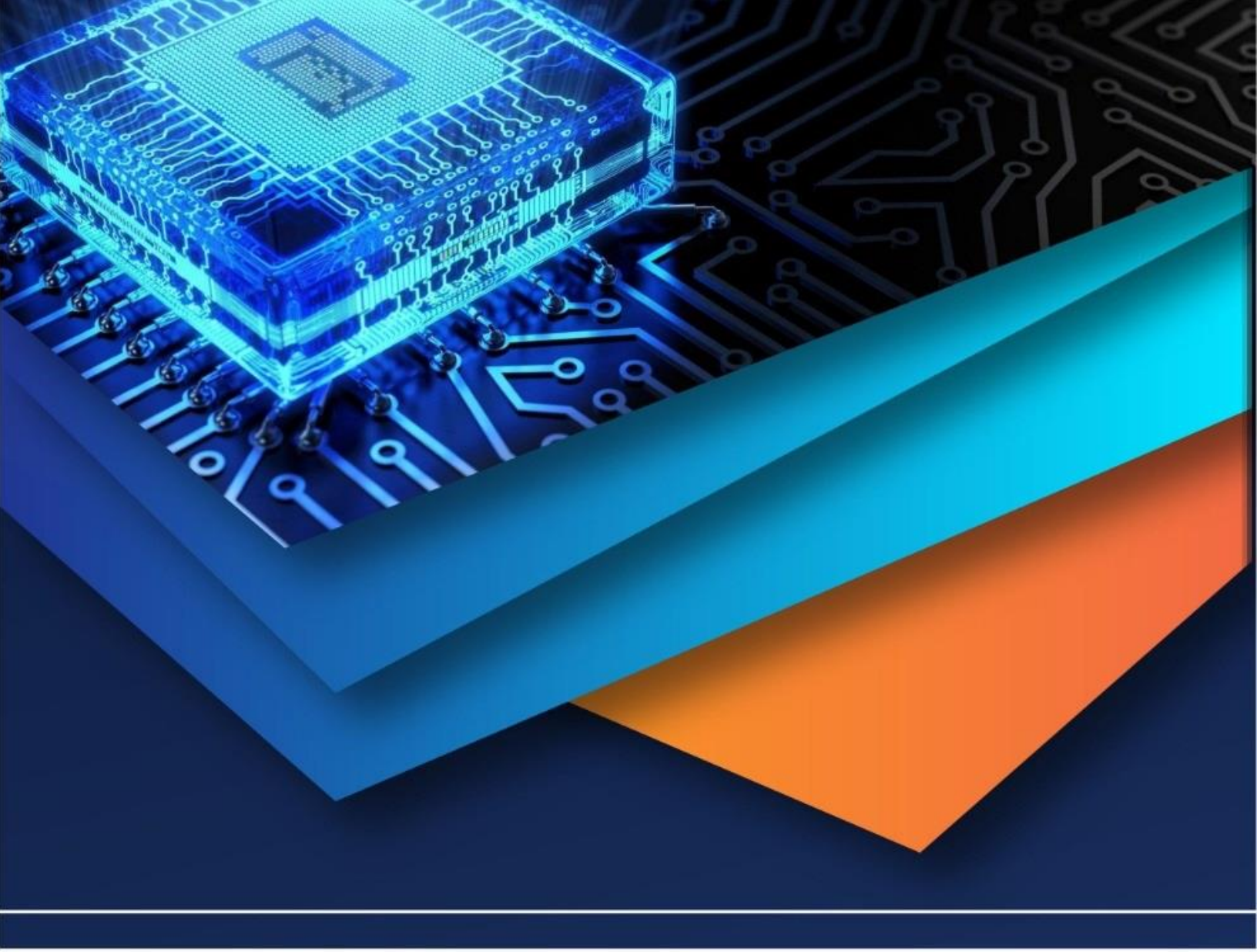
V. CONCLUSION

This paper provides a various classification algorithms of the student prediction for placement training. Future studies will investigate new hybrid models of classification algorithms to improve the performance of prediction system.

REFERENCES

[1] Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students" Performance ", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, Page no. 63-70 2011.

[2] Dr. Abdullah AL-Malaise, Dr. Areej Malibariand Mona Alkhozae," Students' Performance Prediction System Using Multi AgentData Mining Technique", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.5, 2014September.
[3] Kamal Bunkar,Rajessh kumar,Umesh Kumar Singhand Bhupendra Pandya," Data Mining: Prediction for Performance Improvement of Graduate Students using Classification", IEEE, DOI-978-1-4673-1989-8/12, 2012.
[4] S.Venkata Krishna Kumarand S.Padmapriya," An Efficient Recommender System for Predicting Study Track to Students Using Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering, ISSN 2278-1021, vol 3, Issue 9, 2014September.
[5] Dorina Kabakchieva," Predicting Student Performance by Using Data Mining Methods for Classification",Bulgarian Academy of Science CyberneticsandInformation Technologies, ISSN: 1314-4081, Volume 13, No 1, DOI 10.2478, 2013.
[6] Bashir Khan, Malik Sikandar Hayat Khiyal and Muhammad Daud Khattak,"Final Grade Prediction of Secondary School Student using Decision Tree", International Journal of Computer Applications, Volume 115 – No. 21, 2015 April.
[7] G. Naga Raja Prasad and Dr. A. Vinaya Babu," Mining Previous Marks Data to Predict Students Performance in Their Final Year Examinations", International Journal of Engineering Research & Technology, ISSN: 2278-0181, Vol. 2 Issue 2, 2013 February.
[8] Jyoti Namdeo and Naveenkumar Jayakumar," Predicting Students Performance Using Data Mining ", International Journal of Advance Research in Computer Science and Management Studies, ISSN: 2321-7782, Volume 2, Issue 2, 2014February.
[9] Md. Hedayetul Islam Shovonand Mahfuza Haque," Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 2, Issue 7, 2012 July.
[10] Rupali Bhardwaj and Sonia Vatta," Implementation of ID3 Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 6, 2013 June.
[11] Meenakshi Sharma," Data Mining: A Literature Survey", International Journal Emerging Research in Management and Technology, ISSN: 2278-9359 , volume 3,issue 2, 2014february.
[12] Arpit Trivedi," Evaluation of Student Classification Based On Decision Tree", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X , Volume 4, Issue 2, 2014February.
[13] Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer and William F. Punch," Predicting Student Performance: An Application of DataMining Methods With The Educational Web-Based SystemLON-CAPA", 33rd ASEE/IEEE Frontiers in Education Conference, DO-0-7803-7444-4/03, 2003 November 5-8.
[14] Mashael A. Al-Barrak and Muna Al-Razgan." Predicting Students Final GPA Using Decision Trees: A Case Study, International Journal of Information and Education Technology, Vol. 6, No. 7, July 2016.
Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao,"predicting students' performance using ID3 and C4.5 classification algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013.
[15] Priyanka A. Patil,R. V. Mane, " Student Performance Prediction by Pattern Mining Technique", Int.J.Computer Technology & Applications,Vol 5 (2),431-434
[16] Surjeet Kumar Yadav, Saurabh Pal, " Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.
[17] Krina Parmar, Prof. Dineshkumar Vaghela, Dr Priyanka Sharma," performance prediction of students using distributed data mining", IEEE sponsored 2nd international conference on Innovations in Information Embedded and Communication Systems, 978-1-4799-6818-3/15, 2015 .



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY