Finding Factors Responsible for Building Academic Performance Confidence

Sunil Kumar Tiwari[1], Dr. Archi Dubey[2]

[1]Research Scholar, MSMSR Dep't, MATS University, Raipur Chhattisgarh,

[2]Research Supervisor, MSBS Dep't MATS University Raipur Chhattisgarh,

ABSTRACT

In decision making process most of the time a problem comes that where to start on which factor will give larger impact on the resultant factor. Decision tree algorithm provide hierarchical architecture of used attribute which participate in decision making rule. Decision tree algorithm C5.0 with information gain categorize data into classes to which it matches most. The research paper uses decision tree C5.0 algorithm to find the hierarchy of factors and path of factors which build academic performance confidence among the student.

Keywords: *Academic Performance factor,C5.0, Decision tree, rule mining.*

I. INTRODUCTION

Decision tree analytical tool is one of the easiest method for decision making. The characteristic of partitioning the dataset on homogeneous basis is effective. Decision trees are considered for supervised classification. Oxford dictionary defines supervised learning as "the process of classifying something according to shared qualities or characteristic". Decision tree comprises various algorithm i.e.Decision trees form a hierarchy of observed variables on their variable selection criterion.

Educational achievement is one of the primary factor to predict the skill of the individual. A lot of effort has been made by components of education system. But there could be other factor which can affect the performance of the student. Researchers of different discipline are working to find and predict the factors responsible for poor academic performance of the student.

In this paper five variable has been studied for the academic performance. the values obtained in these variables are studied using C5.0 decision tree algorithm to find the hierarchy of the factor responsible for academic performance.

II.LITERATURE REVIEW

Patil N et.al [1] studied 2000 record using c5.0 and CART. The accuracy rate in this research was 99.6 for C5.0 whereas CART has 94.5 accuracy. They applied classification algorithm for customer membership card and opined that children number & income level affects card ranks.

Revathy R and Lawrence R [2] applied C4.5 and C5.0 on crop pest data. On comparison of C4.5 and C5.0 on crop pest data they opined that C5.0 is powerful and preferred method.

Patel B R et.al. [3] organized their study on CHID, CART, ID3, C4.5, C5.0 and Huts algorithm. In the research paper they highlighted major issues related to the decision tree e.g. fragmentation problem, replication problem, partitioning in continuous data, incremental learning problem, handling of range inputs, overfitting etc.

Dai Q et.al. [4] in their research opined that decision tree algorithm generate understandable rules. Other important fact about the decision tree algorithm, they show more important attribute for forming any rule.

Mesaric J et.al. [5] uses decision tree classification algorithm for predicting student success. They opined that it is difficult to predict achievement at one stage, will continue at next stage. Limitation of their study, they highlighted student promoted to next level without passing all first-year courses.

Al-Barrak M A et.al. [6] predicted GPA of students using decision tree algorithm. After analyzing the data of student, they presented list of subjects which is important in each semester for predicting the GPA. In their research work they used J48 decision tree algorithm for the study.

Patel B N [7] presented a comparative summary of various classification techniques. They compared decision tree, neural network, KNN and SVM. In a table they provide comparison among the tools. Table shows that on factor speed of classification and explanation ability decision tree are best. In their research work conclusion, they mentioned that decision tree “provides a theoretical framework for taking into account not only the experimental data to design an optimal classifier, but also a structural behavior for allowing better generalization capability”.

Krishna Kumar S V et.al. chosen 6 factors i.e. type of data, speed, pruning, missing value. Pruning and formula for comparing decision tree algorithm C4.5, CART, ID3 and C5.0. they concluded that C5.0 algorithm is better than other algorithm in their study.

III. C5.0 ALGORITHM

C4.5 and ID3 decision tree algorithm had some limitation. To overcome these limitations a new decision tree algorithm introduced and named C5.0. C5.0 algorithm study multivariate attribute. C5.0 has capability to handle missing values also. The core concept of C5.0 is based on entropy and information gain.

The entropy is defined as

“Let S be a random variable with outcomes Si, i$\in\{1,2\ldots..n\}$ and probability mass function p. The quantity”[9]

$$Entropy = \sum_{i=1}^{n} -p(Si)\log_2(p(Si))$$

“Entropy is used to measure the amount of uncertainty or surprise or randomness in a set of data”[10]

Information gain describe reduction of entropy when a data set split on an attribute into two subsets. Information gain is calculated as

$$Gain\ (S,A) = Entropy\,(S) - \sum_{v\in Values(A)} \frac{|Sv|}{S} Entropy\,(Sv)$$

Where

Values(A)= Set of all possible values

Sv=Subset of S for which attribute A has Value

S=Set of all Values

C5.0 algorithm uses a data set, say D, as a set of training tuple and its associated label of class. Attributes of dataset are used as split pint at different levels of decision tree. C5.0 algorithm uses best splitting criteria on the basis of information gain.

IV. RESEARCH METHODOLOGY

For the research work data is collected as primary data from the target institution and studied with well define step as shown in the figure 1. Data is collected against some question which tends to studied attribute.

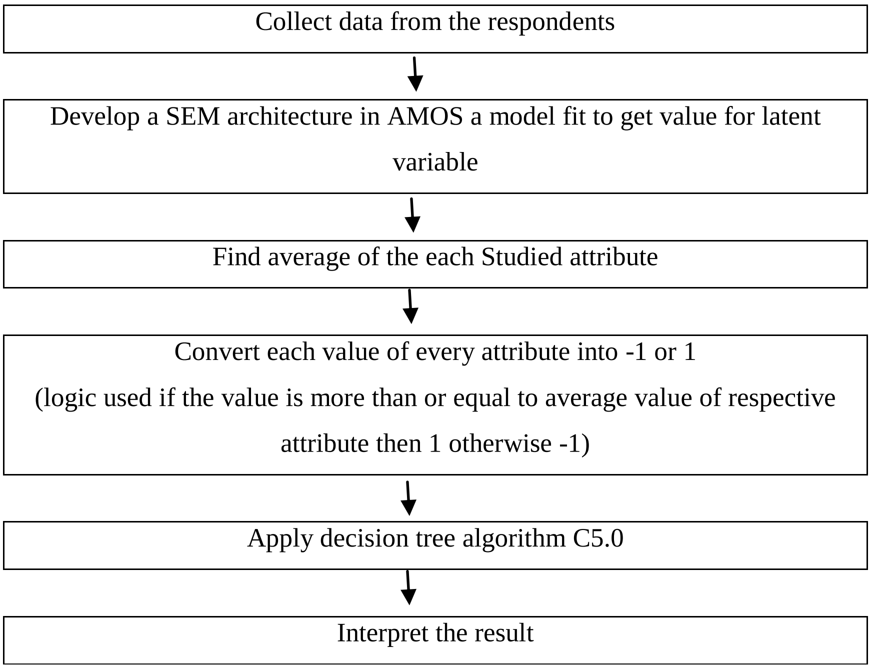
Collect data from the respondents
Develop a SEM architecture in AMOS a model fit to get value for latent variable
Find average of the each Studied attribute
Convert each value of every attribute into -1 or 1
(logic used if the value is more than or equal to average value of respective attribute then 1 otherwise -1)
Apply decision tree algorithm C5.0
Interpret the result

Fig 1: Research Methodology of the work

V. DATA SET

During the month of Nov 2017, primary data collected from the randomly selected student studying at higher secondary school from two District Baloda Bazar (Government Higher Secondary School Lawan, Government managed) and Raipur (Disha college of higher secondary studies, Kota, Raipur, Privately managed) of Chhattisgarh. The survey includes 500 student responses in completing the self-administered questionnaire. Out of these 461 questionnaires were found for the research purpose.

Table 1: Studied factors

Sr. No.	Code	Description
1	SPE	School Physical Environment
2	TSR	Teacher Student Relationship
3	SS	Study Satisfaction
4	LGI	Guardian Interest
5	AP	Academic Performance confidence

In table 1 SPE, TSR, SS, LGI and AP are considered for student performance confidence among the student.

VI. DATA ANALYSIS TOOL

C5.0 algorithm is used to study this data. R programming language has “C50” named package for C5.0. Data is stored in the excel. R has “readxl” package to read the data from excel files. The name of the data file is data2_c.xls.

Table 2: Program code used for study in r

```
)
0)
xl)
 <-     read_excel("C:/Users/USER/Desktop/New     folder/sunil/model
 1/data2_C.xls", sheet = "Sheet2")
50)
ow(sunilc50))
sunilc50[order(g),]
P<-as.factor(sunilc50r$AP)
R","SPE","SS","LGI")
=sunilc50r[,vars], y=sunilc50r$AP, rules = TRUE)
unilc50r$AP ~ TSR +SPE+SS+LGI,data = sunilc50r)



1)
2)
```

VII. RESULT ANALYSIS

C5.0 algorithm used in R language for this study. The code developed in R produces attribute usage, pattern, error rate and confusion matrix of classification. Attribute usage is shown in the table 3. Table 3 indicates that TSR has highest attribute usage whereas SS has lowest usage. The attribute usage indicates that only 3 attributes is used for predicting the class out of 4.

Table 3: Attribute usage in the algorithm

Attribute Usage	**Percentage**
TSR	100.00%
LGI	100.00%
SS	15.84%

Table 4: Rules identified C5.0

```
TSR > -1:
:...LGI <= -1: -1 (116/54)
:   LGI > -1: 1 (150/45)
TSR <= -1:
:...LGI <= -1: -1 (122/22)
    LGI > -1:
:...SS <= -1: -1 (53/20)
        SS > -1: 1 (20/6)
```

From Table 4 and figure 2 it is inferred that

- When TSR is positive then LGI is the only attribute which decide the student academic performance confidence.
- If TSR is positiveand LGI is negative then probability of negative academic performance confidence is high.
- If TSR is negative then LGI attribute is more important to decide the academic performance confidence.
- If TSR is negative and LGI is also negative then probability of negative academic performance confidence is high.
- IF TSR is negative and LGI is positive in that case SS plays importance role to predict the academic performance confidence.
- IF TSR is negative LGI is positive and SS is negative then probability of negative academic performance confidence is high.
- If TSR is negative LGI is positive and SS is positive then the probability of positive academic performance confidence is high.

From the rule identified in table 4 the number of tuple matching to specific class shown in the following table 5:

Table 5: Number of tuples related to the class and their probability

TSR	LGI	SS	1 (Probability)	-1 (Probability)	Total
1	1		105 (0.7)	45 (0.3)	150

	-1		54 (0.465)	62 (0.535)	116
-1	1	1	14 (0.7)	6 (0.3)	20
		-1	33 (0.623)	20 (0.377)	53
	-1		22 (0.18)	100 (0.82)	122

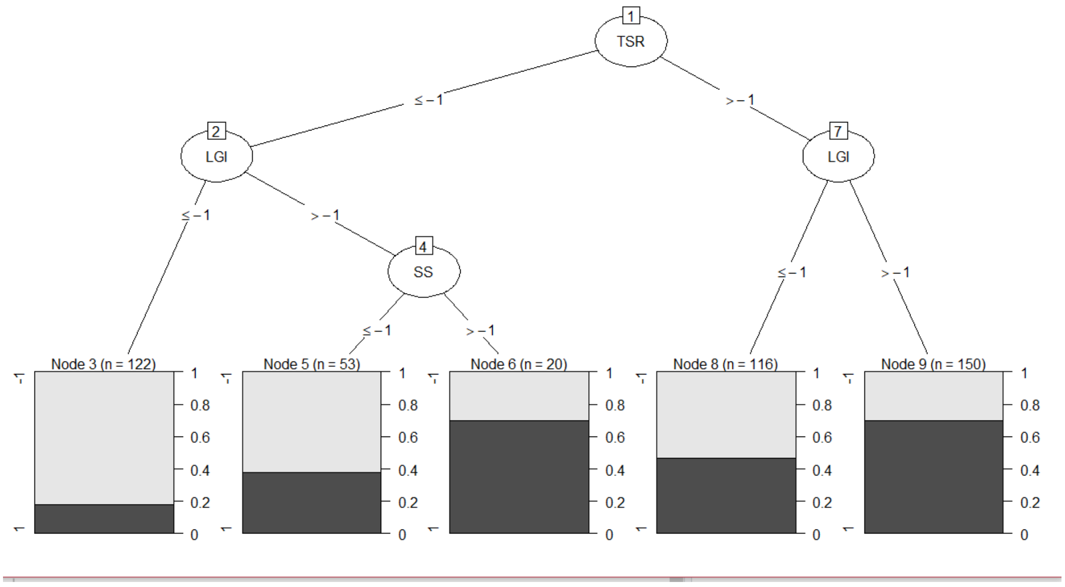

Fig 2: Decision tree chart of studied dataset using C5.0 algorithm

Table 6: Number of tuples allocated to the class and error rate of classification

```
Decision Tree
                  ----------------
                  Size      Errors
                     5  147(31.9%)
                   (a)   (b)    <-classified as
                  ----  ----
                   195    51    (a): class -1
                    96   119    (b): class 1
```

Studied data set of size 461 cases has size of 5 and 31.9% error is prediction. Out of 461 cases 147 are misclassified. Total 246 are in negative class and 215 are in positive class. Out of 246 from negative class 195 are properly classified whereas 51 are misclassified. Out of 215 from positive class 119 are properly classified whereas 96 are misclassified.

VIII.CONCLUSION

Decision tree draws a conditional diagram to show the path among the attribute. The research paper discussed five factors teacher student relationship (TSR), guardian interest (LGI), study satisfaction (SS), school physical environment (SPE) and academic performance confidence (AP). These attributes studied using C5.0 algorithm in R programing to identify the attribute which builds academic confidence among the student. The study reveals that TSR and LGI are dominating factor to build academic performance confidence.

From the study it is concluded that institution must focus on the teacher student relationship as well as build guardian interest among with student interest to retain them in the institution.

REFERENCES

1. Patil, N., Lathi, R.,Chitre, V., “Comparison of C5.0 & CART Classification algorithms using pruning Technique”, International Journal of Engineering Research & Technology (IJERT), Volume.1, Issue.4, June 2012, pp: 1-5.
2. Revathy R, Lawrance R, “Comparative analysis of C4.5 and C5.0 Algorithms on Crop Pest Data”, International Journal of Innovative Research in Computer and Communication Engineering, Vol 5 Special Issue 1 march 2017, ISSN- 2320-9801-page no. 50-58
3. Patel B R & Rana K K, “A Survey on Decision Tree Algorithm For Classification”, IJEDR, Vol 2 Issue 1ISSN 2321-9939 Page No. 1-5
4. Dai Q, Zhang C and Wu H, “Research of Decision Tree Classification in Data Mining”, International Journal of Database Theory and Application, Vol.9 No. 5 (2016) PP.1-8 http://dx.doi.org/10.14257/ijdta.2016.9.5.01
5. Mesaric J and Sebalj D, “Decision tree for predicting the academic success of students”, Croation Operational research Review, CRORR 7(2016), 367-388
6. Al-Barrak M A and Al-Razgan M, “Predicting Students final GPA Using Decision Trees” A Case Study”, International Journal of Information and Education Technology, Vol. 6, No. 7 July 2016 528DOI: 10.7763/IJIET.2016.V6.745 PP 528-533
7. Patel B N , Prajapati S G and Lakhtaria K I, “Efficient Classification of Data Using Decision Tree”, Bonfiring International Journal of Data Mining, Vol. 2 No. 1 March 2012, ISSN 2277-5048
8. Krishna Kumar SV and Kiruthika P, “ An Overview of Classification Algorithm in Data Mining”, International Journal of Advance Research in Computer and Communication Engineering, Vol 4 Issue 12 December 2015, ISSN 2278-1021. Page 255-257.

9. Jansson J, “Decision Tree Classification of Products Using C5.0 and Prediction of workload Using Time Series Analysis”, ExamensarbeteInomElektroteknik, AvanceradNiva, 30 HP Stoclholm, Sverige 2016

10. Dunham M H, “Data Mining Introductory and Advance Topics”, ISBN: 978-81-7758-2 Page 93