DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.4905

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

HEURISTIC METHODS TO IDENTIFY FACTORS INFLUENCING COLLEGE ADMISSIONS IN SUB-URBAN AREAS

Dr. C.Victoria Priscilla
Associate Professor, PG Department of Computer Science
S.D.N.B Vaishnav College for Women, Chromepet,
Chennai, India

Abstract: Now-a-days with growing technology and well connected network the parents and students are evaluating all the information before choosing the right institution. Moreover, there are very few institutions in India who are giving quality inputs and striving to inculcate the learning skills amongst students. With such scenario, the choice of college made by the students and parents are influenced by certain factors like location, adequate infrastructure, sports quota, and much more. Parents and Students in suburban areas have certain expectations about the choice of college. The present article identifies the best key factors that have gained more importance among the parents and students in Kancheepuram district in Tamil Nadu. The dataset taken in the study constitutes of 400 respondents. This paper proposes the Decision tree pruning method to extract the optimal factors that influence the college admissions in suburban areas. Attribute selection measures using Greedy and Ranker method is implemented to prune the decision tree.

Keywords: College admissions, Factors, Decision tree, Pruning, Attribute selection, Greedy method, Ranker method

I. INTRODUCTION

Suburban areas are large residential which surrounds main cities, and urban areas refer to core areas of cities. Suburbs consist mostly of single-family housing far from the city. The lifestyle of people living in suburbs is invariably different from that of the city. So, the expectations of the people vary in all aspects, as in case of housing, education, employment and finance. With such geographical environment, education in suburban areas always has its own demand. An individual learns and gains knowledge only in an educational framework. Nowadays higher education is an important sector for the growth and development of human resource which can take responsibility for economical, social and scientific development of the country [2]. So, colleges in suburban areas have the responsibility to fulfill the requirement of the students living in that area. There are many factors that are expected by the students while choosing the college. Different opinion amongst the parents and students make each factor to carry different priorities in making the choice of college. Priority of students differs with that of their parents. Listing those factors and identifying the factors that are influential in choosing the college is crux of this research. Data mining techniques are used to study the difference in the characteristics between rural and urban students. This kind of problem is quite universal. Present method used WEKA, the open source program for all the computations carried out and the results shown are in the format of WEKA. Literature surveys reveal different factors for different kind of problem.

Kinzie et.al has identified that many factors plays an important part in a student's decision-making process, which includes location near to home, majors offered, costs, help of financial assistance or scholarships offered, selectivity, environment, and parental influence [3]. Shiao-Chuan Kung in his survey of 380 records identified 11 factors which plays significant role in the decision making process [7]. Moore, E. J. et.al have studied economic factors influencing educational attainment and aspirations of farm youth from both rural and urban [4]. Mutekwe et.al studied on career choices and identified that different factors plays significant part on female students in selecting high schools in Zimbabwe [5]. Dubey Pushkar in his finding has identified key factors that contribute to the student's decision in selecting an engineering college in the Odisha[1]. Shammot studied Jordanian students and defined the role of the marketing factors influencing the choice of a private university. The most important factor affected the students choice of the university was the financial costs, and the least factor was the parent's pressure. It was found that males give more attention to the cost than females [6]. The author [8] in her previous study identified about the factors that influence admissions in college according to the parent's and student's perspective views. The author[8] used decision tree to identify the common factors that decided the choice of college among the parents and students.

Now, this article shapes down the research more particular by pruning the decision tree that was developed by the author in her previous study [8]. Attribute selection measures using Greedy and Ranker method is implemented to prune the decision tree. Section 2 describes the data collection method and dataset that will be used for research. Section 3 describes the methodology to identify the best factors. Section 4 interprets the results and discusses the factors that are influential in admissions. Section 5 concludes with future scope of the research.

II. DATASET AND DESCRIPTION

The same dataset that was collected by the author in previous study [8] was used for this research. Survey research method was used to collect the data. Set of different questions related to the students and parents view in choosing the college was framed. Those questions were circulated to the students and parents of suburban areas to know their choice in selection of college. The factors like look of college, goodwill, location of the college, extension of studies, self-finance or government, infrastructure, relationship, placement, favourite course,

alumni, gender, career development, sports, hostel, safety, surrounding of campus were decided as the features of the dataset. 400 instances were collected in and around from Kancheepuram district from both students and parents. The questions with description and option that can be selected are given in Table I.

Table I. Survey Questions with Descriptions

Sr. No	***Question***	*Factors related to questions*
1	Selection of college based on location	Location
2	Selection of college based on government or self-finance	Type
3	Selection of college based on favourite course	Favourite course
4	Selection of college based on infrastructure	Infrastructure
5	Selection of college based on look of college	Look of college
6	Selection of college based on goodwill	Goodwill
7	Selection of college based on alumni	Alumni
8	Selection of college based on placement	Placement
9	Selection of college based on gender	Gender
10	Selection of college based on sports	Sports
11	Selection of college based on hostel	Hostel
12	Selection of college based on extension of studies	Extension of studies
13	Selection of college based on career development	Career development
14	Selection of college based on safety	Safety
15	Selection of college based on surrounding of Campus	Surrounding of campus

The dataset created had the factors as features or attributes or columns. Each record of the dataset is differentiated by the last column which gives the information as whether the recorded response is from the parent's view or student's view. Thus the dataset contained 400 instances with 16 features.

III. METHODOLOGY

Data mining [10] is the process of analyzing data from large data sets on different perspectives and extract meaningful and useful information that can be used to acquire insight into the data. Data mining [9] is used today in diversified applications by researchers and educational institutions for gaining knowledge. In the case of educational institutions and researchers working on problems related to education, they concentrate on the performance analysis of the students with reference to a particular aspect or develop a model that would help the institutions for better performance.

As mentioned earlier, the method is used to prune the decision tree using attribute selection measure. An attribute selection measure is a kind of best search method which selects the best splitting criterion in a given data partition, D, of class-labeled training tuples into individual classes. Attribute measures are also known as "best splitting criterion" because they determine how the tuples at a given node are to be split[11][12]. This paper uses two types of attribute selection measure namely the Greedy method and the Ranker method. This section describes three popular attribute selection measures-attributes subset selection (Greedy method), information gain and gain ratio (Ranker method).

A. Attribute Subset Selection

Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions). The attribute subset selection method determines minimal set of attributes so that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes[11][12].For n attributes, there are 2^n possible subsets. An exhaustive search for the optimal subset of attributes can be prohibitively expensive, especially as n and the number of data classes increase [11] [12]. The "best" attributes are found using statistical test for significance that assumes the attributes are independent of one another. "These methods are greedy in nature while estimating an optimal solution. Mining on a reduced set of attributes has an additional benefit where based on the discovered patterns, the method reduces the number of features to make the patterns easier to understand. From the dataset, the best optimal attributes are selected and the Decision tree has been pruned which is shown in the Table II and in the Figure 1.

Table II. Selection of Factors using Attribute Subset method

Sr. No.	***Best Attribute Subset***
1.	Location
2.	Type
3.	Infrastructure
4	Good will
5.	Alumni

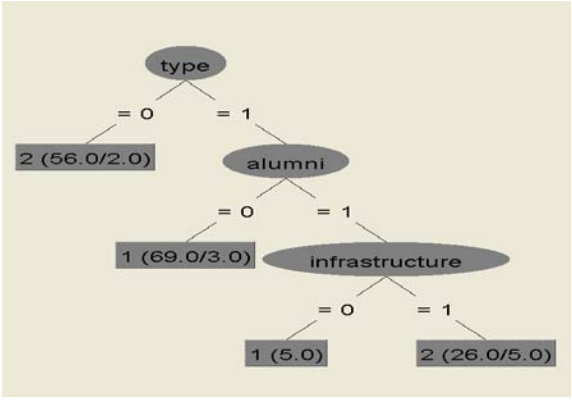

Figure 1. Pruned Tree Based on Best Splitting

The factors are identified as location, goodwill, type, alumni and infrastructure.

Table III. Factors Selected using Ranking Value Based on Information Gain

B. Information Gain

Information gain is defined as a ranking method of attribute

Sr. No.	*Entropy Value*	*Attribute Name*	*Sr. No.*	*Entropy Value*	*Attribute Name*
1.	0.41056	Type	9.	0.02184	Career development
2.	0.19517	Alumni	10.	0.0191	goodwill
3.	0.13845	Infrastructure	11.	0.00919	Favourite course
4.	0.08958	Sports	12.	0.00812	Safety
5.	0.08814	Extension of studies	13.	0.007	Surrounding of campus
6.	0.06946	Hostel facilities	14.	0.00492	Gender
7.	0.04395	Location	15.	0.00412	placement
8.	0.03133	Look of college			

selection measure that is based on information theory, which studied the value or "information content" of Data [11][12]. The node N represents the tuples of partition D and the attribute with the highest information gain is chosen as the splitting attribute for node N [11][12]. This attribute minimizes the information that is needed to classify the instances which is in the resulting partitions and reflects the least randomness or "impurity" in these partitions [11][12]. This kind of approach minimizes the expected number of tests needed to classify a given instance and guarantees that a simple tree is found. The expected information [11][12] needed to classify an instance in D is given by

$$\text{Info}(D) = \sum_{i=1}^{-m} p_i \log_2(p_i) \qquad (1)$$

where p_i is the probability that an arbitrary instance in D belongs to class C_i and is estimated by |Ci,D|/|D|.A log function to the base 2 is used, because the information is encoded in bits. Info(D) is just the average amount of information needed to identify the class label of a instance in D. The information is based on the proportions of instances of each class. Info(D) is also known as the entropy of D[11][12]. This amount of purity for the Attribute (A) is measured by

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \; x \; \text{info}(D) \qquad (2)$$

The term $|D_j| / |D|$ is the weight of the j^{th} partition. Info_A (D) is the expected information to classify a tuple from D based on the partitioning by A. The smaller the expected information required, the greater the purity of the partitions [11] [12].

Information gain [11][12] is defined as the difference between the original information requirements

$$\text{Gain}(A) = \text{Info}(D) - \text{InfoA}(D). \qquad (3)$$

Hence the ranking of each attribute represent in the Table III and the tree has pruned based on the entropy value greater than 0.05 which is shown in the Figure 2

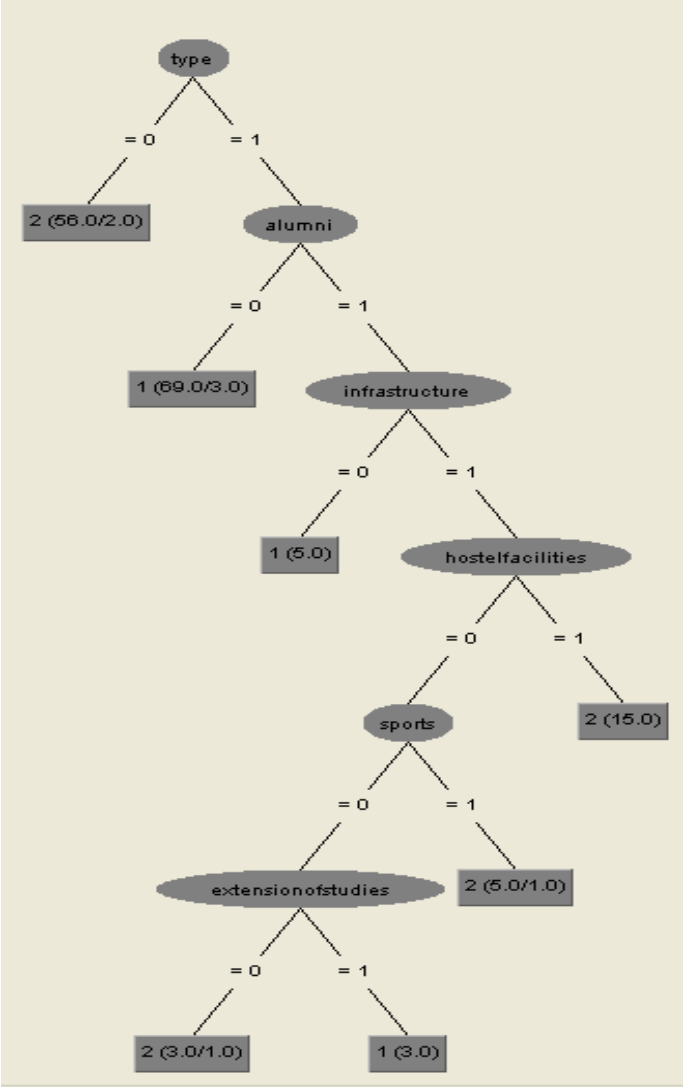

Figure 2. Pruned Tree Based on Information Gain

The factors identified in Information Gain ratio method are type, alumni, infrastructure, hostel facilities, sports and extension of studies.

C. Gain Ratio

Gain Ratio is another ranking method which is used for best attribute selection. It applies a kind of normalization to information gain using a "split information" value defined as

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} x \log_2\left(\frac{D_j}{D}\right) \qquad (4)$$

Split Information value represents the potential information generated by splitting the training data set D, into v partitions, corresponding to the v outcomes of a test on attribute A. For each outcome, it is noted that the method considers the number of tuples having that outcome with

respect to the total number of tuples in D. Based on the same partitioning , information gain measures the information with respect to classification that is acquired [11][12]. The gain ratio is defined as

$$GainRatio(A) = Gain(A) / SplitInfo(A) \quad (5)$$

The attribute with the maximum gain ratio is selected as the splitting attribute which is shown in the Table IV and the tree has pruned based on the gain ratio value greater than 0.05 and shown in the Figure 3.

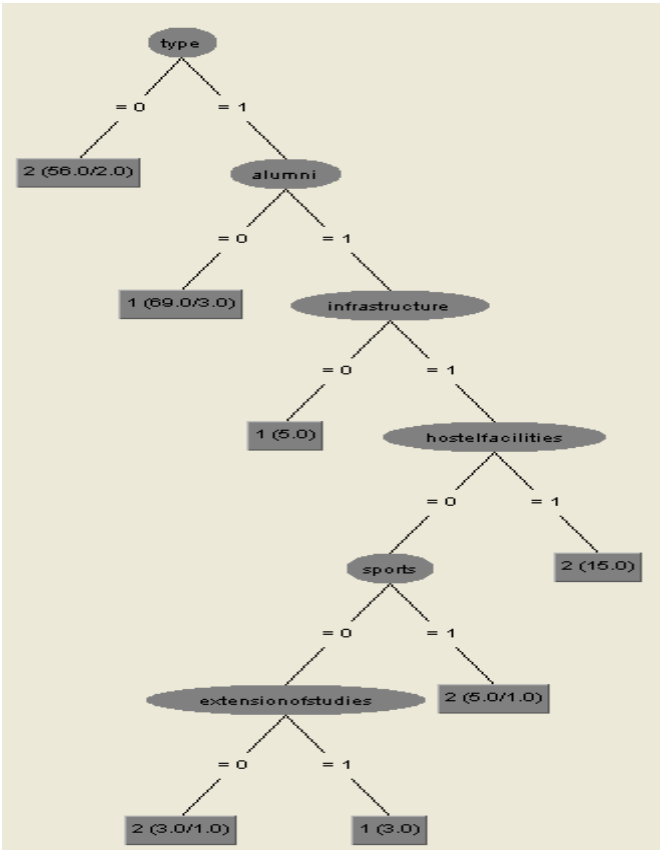

Figure 3. Pruned Tree based on Gain Ratio

Table IV. Factors selected using Ranking Value Based on Gain Ratio

Sr. No.	*Entropy Value*	*Attribute Name*	*Sr. No.*	*Entropy Value*	*Attribute Name*
1.	0.43592	Type	9.	0.0349	Look of college
2.	0.20501	Alumni	10.	0.02934	Career development
3.	0.18912	Infrastructure	11.	0.02363	Safety
4.	0.09258	Extension of studies	12.	0.0131	Surrounding of campus
5.	0.09045	Sports	13.	0.01277	Favourite course
6.	0.07229	Goodwill	14.	0.0112	Placement
7.	0.0695	Hostel facilities	15.	0.0071	Gender
8.	0.0564	Location			

The factors identified in Gain ratio method are type, alumni, infrastructure, hostel facilities, sports and extension of studies.

IV. Interpretation of Results and Discussion

The problem addressed in this research is to identify the factors that influence the college admissions in sub-urban areas. The previous findings of the author [8] identified the factors as Alumni, Infrastructure, Hostel, Sports and Look of college. The key factors that were identified in this paper using the greedy method and the ranker method are compared with the previous study and are tabulated in Table V. From Table V it is ascertained that the optimal key factors that influence the college admissions in sub-urban areas are Type, Alumni, Infrastructure, Hostel and Sports.

Table V. Comparison analysis for finding common Key

Method	*Factors*
Greedy method – Attribute subset selection	Location, Goodwill, Type, Alumni, Infrastructure
Ranker method – Information Gain	Type, Alumni, Infrastructure, Hostel, sports, extension of studies
Ranker method- Gain Ratio	Type, Alumni, Infrastructure, Hostel, sports, extension of studies.
Decision tree [8]	Alumni, Infrastructure, Hostel, Sports, Look of college

V. Conclusion and Future Work

The study was implemented for the data that was collected within the sub urban areas. The data consisted of both parents and students view in the choice of college and using different decision tree pruning, the factors were identified as Type, Alumni, Infrastructure, Hostel and Sports. The model implemented in this paper is easy to be read and understood. This model can give the interesting information about the admissions in suburban area and provides guidance to students to make the decision. The study can be enhanced in future by collecting data from the city area and compared with that of this sub-urban area.

VI. References

[1] Dubey Pushkar, Sharma Sudhir Kumar, Surenthiran N, "Factors affecting choice of engineering colleges in odisha" ,Research Journal of Management Sciences, 2013.

[2] Kanunjna D., Higher education in India: Some relevant issues and concerns. University News., 50(14), pp.1-4, 2012.

[3] Kinzie, J., Palmer, M., Hayek, D., Hossler, D., Jacob, S. A., & Cummings, H,. Fifty years of college choice: Social, political and institutional influences on the decision-making Process. Indianapolis, IN: Lumina Foundation for Education., 2004.

[4] Moore, E. J., Baum, E. L., & Glasgow,R. B. , Economic factors influencing educational attainment and aspirations of farm youth. Washington, DC:Economic Research Service, Resource Development Division. (ERIC Document Reproduction Service Document No. ED015797), pp.1-43. 1984.

[5] Mutekwe E., Modiba M. and Maphosa M., Factors Affecting Female Students' Career Choices and Aspirations: A Zimbabwean Example., J Soc Sci, 292), pp.133-141 2011.

[6] Shammot M.M., Factors Affecting the Jordanian Students' Selection Decision Among Private Universities Journal of Business Studies Quarterly.,2(3), pp.57-63,2011.

[7] Shiao-Chuan Kung., Factors that Affect Students' Decision to Take distance Learning Courses: A Survey Study of Technical College Students in Taiwan., International Council for Education Media., pp.299-305, 2002

[8] C. Victoria Priscilla, M.Mahadevi , "A Parent-Student Perspective View in Selecting the College in Suburban Area", International Journal of Computer Applications (0975 – 8887) 166(10), May 2017.

[9] R.Yogatharani, "A Study on Classifier Performance Using Machine Learning Algorithm", International Journal of Scientific Engineering and Applied Science (IJSEAS) –1(3), 2015.

[10] Margret H. Dunham, "Data Mining: Introductory and advance topic".

[11] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann.

[12] http://hanj.cs.illinois.edu/cs412/bk3/08.pdf