International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 5 ǁ May 2023 ǁ PP. 599-605

Deep Mining of Factors Influencing the Ability of College Students in a Smart Campus

QIAOYi-fan
School of Science and Technology, Tianjin University of Finance and Economics,Tianjin City, China
Corresponding Author:QIAOYi-fan

Abstract
In the context of the comprehensive improvement of social requirements for college students' abilities, how to deeply mining the influencing factors of college students' abilities has become a difficult problem that both students and universities need to solve. On the basis of the massive data accumulated in smart campuses, in response to various factors such as individuals, families, and universities that affect the development of college students' abilities, K-means algorithm is first used to cluster and analyze college students' abilities. Then, Apriori algorithm is used to deeply mining the influencing factors of college students' abilities and identify the main factors that affect their abilities. The research results on the one hand point out the direction for college students to improve their abilities, On the other hand, it also provides reference for the policy-making of higher education and teaching.

Keywords:*College students' abilities, K-means clustering, Apriori algorithm*

Date of Submission: 09-05-2023 Date of acceptance: 20-05-2023

I. INTRODUCTION

With the increasing number of college graduates year by year and the arrival of the most difficult employment season every year, society's requirements for college students' abilities are no longer limited to their grades, but have put forward higher requirements for their abilities. For example, recruiting graduate students in universities requires students to have outstanding learning abilities, research institutions require students to have strong research abilities, enterprise technical positions require students to have good hands-on skills, enterprise management positions require students to have strong communication and management skills, and even some units have high requirements for multiple abilities in universities.

The ability development of college students during their school years belongs to a process of autonomous development. Originally, the main focus was on the internal influencing factors of students and their internal changes, while universities paid less attention to the impact of college students' ability development. However, as an extremely important part of the growth of college students, the impact of universities cannot be ignored. The impact model of universitiesstarts with the factors that affect the development of college students' abilities[1], pays attention to the improvement of students' abilities during their college years and the impact of the university environment on their ability development, and connects college students' abilities with their families, individuals, and the university environment.

How to improve the various abilities of college students has become a difficult problem that both students and universities need to solve. Therefore, studying the influencing factors of college students' abilities not only meets social requirements but also conforms to the development of the times. At the same time, the introduction of smart campuses has accumulated a massive amount of underutilized data, providing a data foundation for quantitative research, discovering and utilizing valuable information to deeply explore the influencing factors of college students' abilities.

II. ANALYSIS OF FACTORS INFLUENCING THE ABILITY OF COLLEGE STUDENTS IN THE IMPACT MODEL OF COLLEGES AND UNIVERSITIES

The institutional impact model consists of three parts: input, environment, and output. The influencing factors on the cultivation of college students' abilities are concentrated in the first two parts: input and environment, and at the output end, the main indicators are various abilities of college students, as shown in Figure 1.

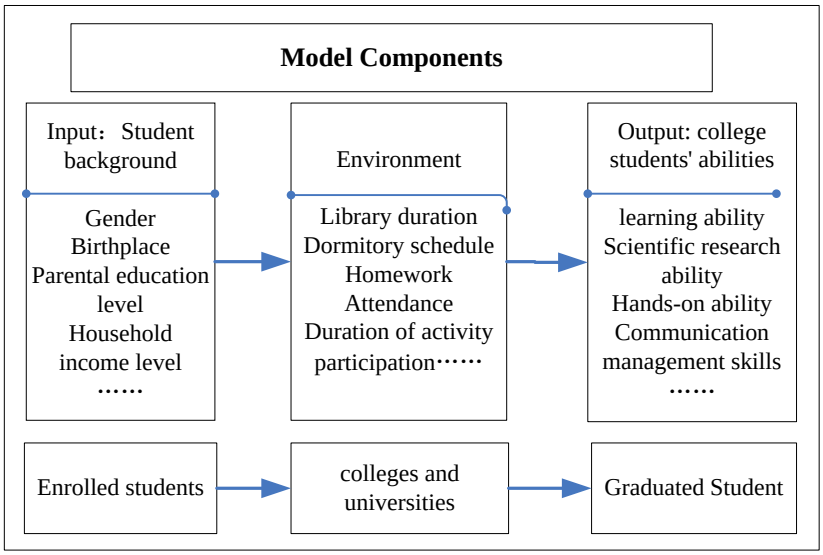

Figure1: Institutional Impact Model

2.1 INPUT

As an input variable in the impact model of universities, students' personal and family backgrounds cannot be ignored in their ability development. According to the analysis of factors in reference 4 and the data accumulated on smart campuses, the personal information statistics table for college enrollment includes factors such as gender, place of origin, parental education level, and family income level, all of which have a certain impact on the personal ability development of college students.

2.2 ENVIRONMENT

As an intermediate processing end of the impact model of universities, universities play an extremely important role in cultivating the abilities of college students. According to the analysis of the impact in reference 1 and the data accumulated on smart campuses, there are many factors that affect the personal development of college students during their study, including library duration, dormitory schedule, teacher level, classroom environment, teaching support, curriculum system, classmates, roommates, and extracurricular activities.

2.3 OUTPUT

After being trained in universities, college students have improved their abilities in various aspects. Referring to all social standards for talent selection, one is their level of knowledge mastery, which is their academic performance in the school; The second is scientific research ability, which refers to the hosting and participation of various scientific research projects; The third is hands-on ability, which directly reflects the participation in various competitions; The fourth is communication and management skills, organizing and participating in various activities, and whether they have served as class or university cadres; The fifth is other abilities, which society requires in all aspects.

III. DATA SET

3.1 DATA SELECTION

Data selection is the first step in data mining and directly affects the results of data mining. Faced with the massive data accumulated in a smart campus, based on the data needs of each part of the impact model of the university, the research object is selected as all students of a certain university in 2017 and 2018, including the basic student information table at the input end; Dormitory allocation table, dormitory schedule, library hours statistics table, attendance statistics table, daily homework statistics table, student schedule, and GPA of the university environment; The student research score table, student competition score table, student management table, etc. at the output end.

3.2 DATA CLEANSINGT

Selecting incomplete, missing, or duplicate data requires data cleaning to eliminate it. For example, in the research points table, if a student's project is not completed properly, this part of the points needs to be removed. In the average GPA table, some students' academic progress in university is no longer consistent with other students due to reasons such as dropping out of school or joining the military, resulting in distorted GPA data. These records also need to be deleted. In order to ensure the accuracy of the calculation, the library hours

and dormitory schedules need to exclude irregular schedules such as weekends, holidays, and during the pandemic.

3.3 DATA INTEGRATION

Summarize and calculate the cleared relevant data, summarize and sum the four year related data of each student based on their research points, competition points, management points, and other related tables, and then integrate the output data with the student number as the primary key according to the school impact model to generate a college student ability score table. Integrate the data from the input end and the university environment with the student ID as the primary key to generate a table of influencing factors for college students.

3.4 DATA CONVERSION

In order to meet the requirements of the algorithm for data, it is necessary to convert some of the original data, such as the gender of students, taking 1 and 0 respectively, the source of students taking the postgraduate enrollment partition AB as 1 and 0 respectively, the parental education level taking higher education as the boundary, taking 1 as higher education, 0 as no, etc.

3.5 DATA REDUCTION

K-means clustering analysis is used for the ability score table of college students. The conversion of student research, competition, and management scores is based on the values specified in the school's comprehensive quality test and will not be adjusted. When using the Apriori algorithm to mine the influencing factors of association rules for the college students influencing factors table, some values in the college students influencing factors table are continuous values, such as the length of the library, the work and rest of the dormitory, and the data form is required by the algorithm. All kinds of data need to be discretization. The original column values are 1 and 0, which remain unchanged. Other columns take their column average as the standard. The column average value greater than or equal to is 1, and the column average value less than is 0. Unrelated attributes such as student names and student numbers that are not related to data mining can be deleted during mining.

IV. CLUSTERING ANALYSIS OF COLLEGE STUDENTS' ABILITIES BASED ON K-MEANS

K-means clustering [2] divides college students into several clusters based on their various abilities, and the clustering results satisfy that students in the same cluster have higher similarity, while students in different clusters have lower similarity. Firstly, preprocess the various abilities of college students and connect them with their student numbers to generate a college student ability score table, as shown in Table 1. Then, use the K-means clustering method to group and cluster the students in the student ability score table 1.

Table 1:Ability Points for College Students.

Student ID	Name	Average GPA	Scientific research points	Competition points	Manage Points
2017*****	Chen**	2.935	109	76	85
2017*****	Song**	3.147	134	152	135
2017*****	Wang**	3.546	240	238	215
2017*****	Wei**	3.182	138	166	150
2017*****	Xu**	3.191	144	184	164

4.1 SELECTION OF K-VALUE FOR THE NUMBER OF CLUSTERS

After preprocessing, the college student ability score table uses the evaluation function score of the sklearn library to test the impact of different k-values on the evaluation indicators, and then selects the appropriate number of clusters. The specific implementation steps are as follows:

i. Data reading, reading the preprocessed college student ability score table 1.
ii. Data normalization processing, removing the meaningless attributes of student ID and name columns, and using the scale function of the preprocessing module in the sklearn library for data normalization processing.
iii. Cluster evaluation uses the range method to generate the number of clusters with K values ranging from 1 to 20. The K-means module in the sklearn library is used to construct a cluster, and different K values are used for clustering training. Cluster evaluation indicators are generated through the score evaluation function.

iv. The results show that a line chart of the corresponding relationship between the number of clusters and the cluster evaluation indicators is generated, as shown in Figure 2.

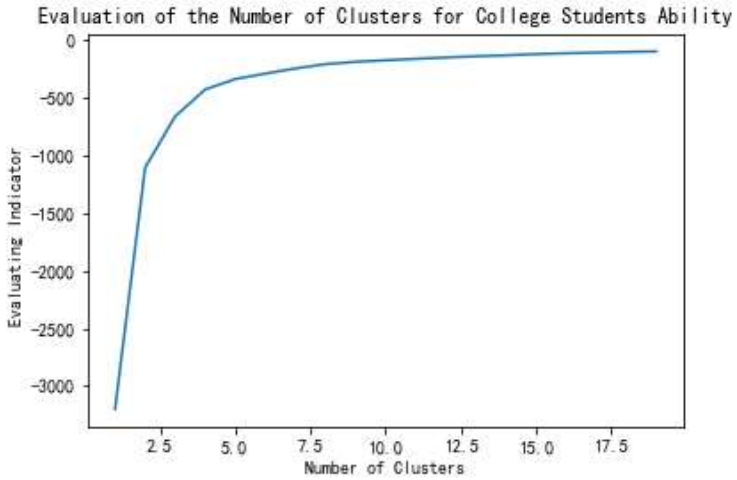

Figure 2: Evaluation of Cluster Number Indicator

From the above figure, it can be seen that when the number of clusters K is greater than 5, the clustering evaluation index tends to stabilize, so the number of clusters K is taken as 5.

4.2 CLUSTER ANALYSIS OF COLLEGE STUDENTS' ABILITIES

After determining the number of clusters, cluster analysis is conducted on the abilities of college students, counting the number of students and cluster centers in each cluster, and drawing the cluster results. The specific implementation steps are as follows:

i. Cluster modeling, constructing a cluster with a K value of 5, and then using the K-means module fit method in the sklearn library for clustering training.

ii. Count the number of students of various types, using the generated kmeans.labels_, Calculate the number of students in each cluster, as shown in the student number column in Table 2. There are more students in the middle and less students at both ends, which conforms to normal distribution.

iii. Identify the cluster center and utilize the generated kmeans.cluster_ centers_, Identify the cluster center values for each capability, as shown in Table 2.

Table 2: Ability Points for College Students

Cluster	Number of students	Average GPA	Scientific research points	Competition points	Manage Points
Cluster0	205	2.910102	84.921951	89.843902	89.746341
Cluster1	212	3.205538	161.797170	149.957547	150.542453
Cluster2	96	3.646990	252.750000	265.781250	265.718750
Cluster3	166	3.377898	193.084337	207.542169	207.283133
Cluster4	121	2.348537	51.553719	41.644628	40.652893

iv. Clustering results, use the sklearn database decomposition module to reduce dimensions to generate a clustering effect diagram. From the diagram, it can be seen that Cluster4 and Cluster2 are the worst and best in terms of various abilities. Cluster0, Cluster1 and Cluster3 are students with moderate abilities. After discretization of college students' abilities, it is convenient to further analyze the factors of various students, as shown in Figure 3.

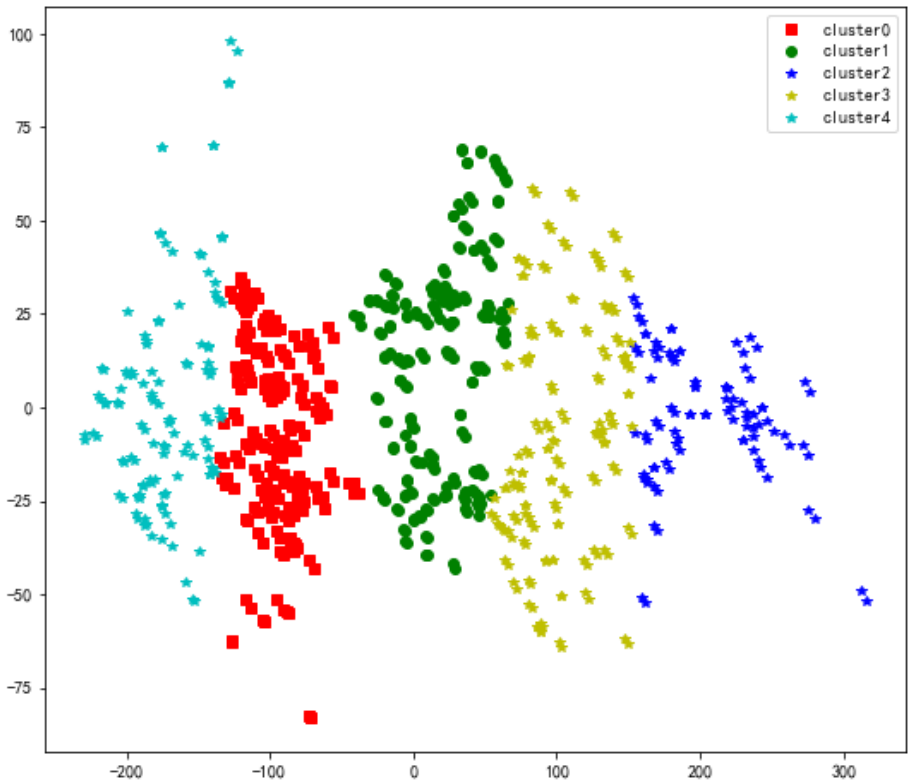

Figure 3: Cluster Results of College Students' Abilities

V. DEEP MINING OF FACTORS INFLUENCING COLLEGE STUDENTS' ABILITY BASED ON APRIORI ALGORITHM

The framework for mining and analyzing the influencing factors of college students' abilities is shown in Figure 4. The processing process is mainly divided into four steps, specifically: collecting data on various ability indicators of college students, clustering analysis of college students' abilities, preprocessing of influencing factors, and mining association rules. After clustering the abilities of college students using the K-means algorithm, in order to further explore the influencing factors, the Apriori algorithm [3] is used to mine association rules for the influencing factors of college students.

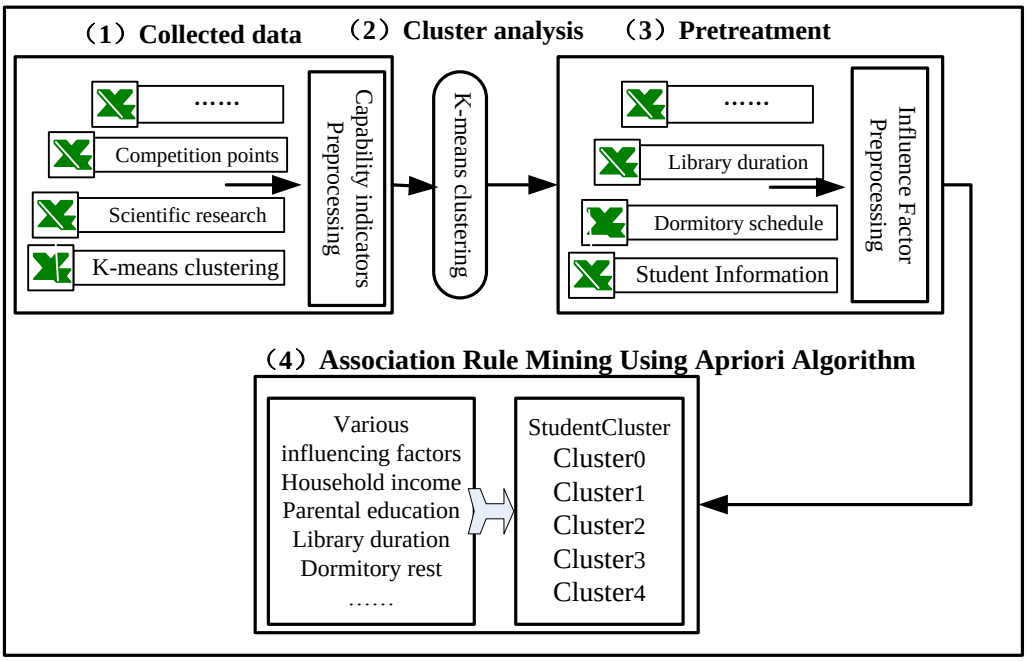

Figure 4: Analysis framework based on k-means clustering and Apriori algorithm

5.1 APRIORI ALGORITHM AND ITS MINING PROCESS

Association rules represent the strong relationship that may exist between two event items, which is called strong association relationship. It is a non supervised learning algorithm. Assuming U is the set of all items in the database, where items A and B belong to U and are independent of each other, S is the set value of all transactions, and $S_{A\to B}$ represents the set value of transactions where item AB occurs simultaneously. Therefore, whether A → B constitutes a strong association rule mainly depends on the support and confidence level set. Among them, Support represents the ratio of the number of simultaneous occurrences of item A and item B in the transaction set to the total transaction set S, expressed as formula (1):

$$\text{Suppport}(\text{A} \to \text{B}) = \frac{S_{A\to B}}{S} \quad (1)$$

Confidence is the number of simultaneous occurrences of item AB, the ratio of which to the number of occurrences of item A, which is the conditional probability. The formula of confidence degree of A to B is expressed as (2):

$$\text{Confidence}(\text{A} \to \text{B}) = \frac{S_{A\to B}}{S_A} \quad (2)$$

The mining process based on the Apriori algorithm involves generating frequent itemsets and strong association rules based on the set minimum support (min_Sup) and minimum confidence (min_Conf), with the following two steps:

(1) Generate frequent itemsets, preprocess a table of influencing factors for college students, scan all transactions one by one, generate candidate options, set the minimum support (min_Sup), and then based on Support ($A \to B$) $\geqslant$ min_ Sup filter candidate, found to meet min_ All itemsets of Sup, also known as frequent itemsets.

(2) Generate strong association rules, utilize frequent itemsets, and set min_ Conf, then Confidence ($A \to B$) $\geqslant$ min_ Conf condition, scan all transactions one by one, find all that meet min_ The association of Conf's rules, also known as strong association rules.

For the influencing factors table of college students after discretization, group student clustering is added with the primary key of student number, the minimum support (min_Sup) and minimum confidence (min_Conf) of association rules are initially set to extract rules, and the threshold is adjusted after multiple experiments until a satisfactory result is extracted, and the association rules that clearly do not have causality are eliminated. When min_ Sup=0.19，min_ When Conf=0.45, satisfactory results can be obtained, as shown in Table 3. For the first time, only the association rules about Cluster3 or Cluster1 are extracted. Because the ability of college students is normally distributed, the number of intermediate students is good. In order to further analyze the influencing factors of the best and worst students in the ability development of college students, the influencing factors of college students are screened again, and the events of Cluster4 (the worst) and Cluster2 (the best) are extracted. Then, the in-depth mining is conducted again. When min_ Sup=0.45，min_ When Conf=0.58, satisfactory results can be obtained, as shown in Table 4.

Table 3: Association Rules for Factors Influencing College Students' Ability.

Related items	Confidence level (%)
Assignments→Cluster3	56.35
Attendance→Cluster3	55.24
Assignments→Cluster1	53.86
Attendance→Cluster1	50.61
Assignments,Attendance→Cluster1	48.47

Table 4:Deep Mining of Factors Influencing College Students' Ability.

Related items	Confidence level(%)
Assignments→Cluster2	95.27
Attendance→Cluster2	93.56
Assignments,Attendance→Cluster2	90.26
Library duration→Cluster2	88.16
Dormitory schedule→Cluster2	85.24
Parental education level	62.14
Household income	56.63

5.2 RESULT ANALYSIS

The development of college students' abilities is influenced by many factors such as individuals, families, and universities. From the analysis results in Table 3, it can be seen that the strong correlation between homework and attendance on college students' abilities indicates that following the arrangements of the school and teachers first in school can help improve college students' abilities, and the relevant requirements of the

school and teachers have a certain degree of scientificity. The impact of factors such as library duration and family income has not been explored. The analysis reason is that with the participation of all students, the set values of minimum support and minimum confidence affect the mining results. However, if these two values are set too low, the mining results will deviate.

In order to further explore the influencing factors of college students' abilities, as the proportion of the two types of students with the worst and best abilities is relatively small, a deep mining was conducted on the college influencing factors table after screening, and Table 4 was obtained. From the analysis results of Table 4, it can be seen that college students have higher confidence in the library time, dormitory work and rest, daily homework, and attendance in universities, with the library time having the highest confidence, It indicates that it is easier to achieve good grades while fully utilizing the school library, and the regularity of dormitory work and rest can also improve the abilities of college students, but it is weaker than the influence of library time. The above two points indicate that students' self-discipline is beneficial for the development of college students. Family income and parental education level have a certain impact on the development of college students' abilities, similar to the study in reference 4.

VI. CONCLUSION

There are many factors that affect the abilities of college students. Based on the massive data of smart campuses, KMeans is used to cluster college students, and Apriori algorithm is used to mine association rules. The main influencing factors of college students' abilities are deeply explored and quantitatively analyzed to identify the main factors that affect their abilities. The research results on the one hand point out the direction for college students to improve their abilities, On the other hand, it also provides reference for the policy-making of higher education and teaching. Thus providing reference opinions for optimizing school teaching. This method is also applicable to other disciplines and has generalizability and applicability.

ACKNOWLEDGMENTS

Funded by the Innovation and Entrepreneurship Training Program for College Students of Tianjin University of Finance andEconomics(202210070257).

REFERENCES

[1]. LU Gen-shu, LIU Xiu-ying. “On College Student Competence Development and Its InfluuencingFactors”J. Journal of Higher Education.2017 Vol.38 No.8, pp.60–68.

[2]. GUO Peng, CAI Cheng. “Data mining and analysis of students’score based on clustering and association algorithm”J. Computer Engineering and Applications, 2019, 55(17):169-179.

[3]. CHEN Ying,CHI Yao-dan,WU BO-qi,LIU An-qi. “Study on Student Curriculum Relevance Based on AprioriAlgorithm”J.Journal of Jilin Jianzhu University, 2019 37(06),64-68.

[4]. Zhou Hai-tao, Jing An-lei, Li Shu-guang. “An Empirical Study of the General Capability of University Students and Its Determinants”J. Journal of Soochow University(Educational Science Edition), 2013,1(01),53-59.