Journal of Education and Educational Research
ISSN: 2957-9465 | Vol. 4, No. 3, 2023

Research on the Impact of AI on University Student Learning based on Principal Component Analysis and Factor Analysis Methods

Yuqi Wang †, Dandan Zhou †

School of Physics and Materials Engineering, Hefei Normal University, Hefei, China

† These authors also contributed equally to this work

Abstract: Artificial intelligence (AI) was introduced in 1956 by scientists from McCarthy and Minsky held at Dartmouth College in the United States. Its rapid development has a different depth of impact on all social strata, but also has a great impact on the education industry, especially for the future college students of the motherland. This paper uses the formal statistical data of the questionnaire to analyze and establish the mathematical model to understand the influence of artificial intelligence on college students in different aspects. First, the raw data is visualized and numericized. Secondly, the analysis results of the data are taken as the evaluation index and comprehensively considered from many aspects such as priority, scientificity and operability. We used principal component analysis and factor analysis to measure each index reasonably. Finally, a comprehensive evaluation method is used to evaluate the influence of AI on college students' learning.

Keywords: Artificial Intelligence; Principal Component Analysis; Factor Analysis.

1. Introduction

In 2016, AlphaGo won a 4-1 victory in an ai-human Go match, showing its outstanding ability in the field of artificial intelligence. Atlas, Boston Dynamics' humanoid robot, also demonstrated its outstanding perception and control. In 2022, the AI painting "Space Opera House" won the first prize in the art competition of the Colorado State Fair, which further proves the huge potential of artificial intelligence in the art field [1]. This paper is in view of the influence of AI on various directions of college students, and respectively expounds the impact of positive and negative aspects, as well as the development, application and views of AI in the future.

2. Data Visualization

In this paper, in order to understand the impact of AI on college students' learning in different aspects, we designed a questionnaire and processed the results by using particle computing methods in analytical classification techniques, using the introduction of data fusion and feature selection and feature extraction of quantitative data, and finally obtained the following visual analysis results (partially shown).

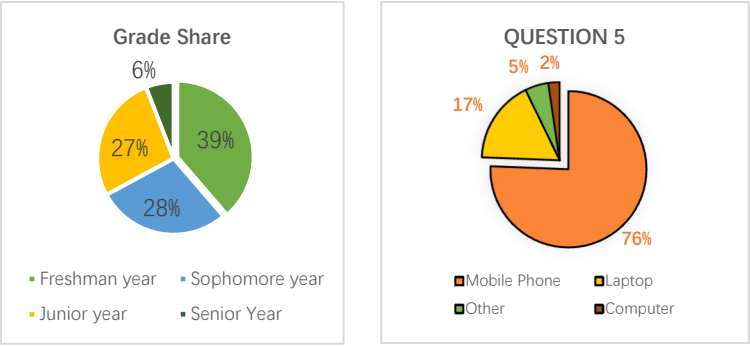


Figure 1. Visualization results

From the visual results, it can be found that among the respondents, the freshman proportion is the most, the fourth year is the least; the most students use mobile phones to surf the Internet.

2.1. Visualizing the Data Analysis Method

2.1.1. Reliability Analysis

Reliability is the main basis of the credibility degree and the reliability of the questionnaire. According to the characteristics and specific requirements of the survey, test the consistency of the scores of each problem in the scale, using the Cronbach coefficient (intrinsic consistency coefficient) [2]. Here, the statistical Cronbach coefficient (α coefficient) is selected as the reliability coefficient for reliability analysis. In most cases, if the coefficient value is greater than 0.8, the tested scale data is very reliable; the Cronbach coefficient above 0.7 can be considered as an acceptable standard. The reliability test using SPSS yielded the following results: the overall Cronbach coefficient α is 0.892, greater than 0.8, and the overall reliability of the questionnaire is good.

2.1.2. Analysis of Validity

Validity refers to the precision of the measurement tool used to test, the degree of its measurement characteristics reflects its measurement effect, and when the validity is

higher, it agrees with the goal of the questionnaire test, so the Bartlett sphericity test and KMO samples were selected.

This topic aims to explore the rationality of the theory and method proposed in this topic. Its effectiveness was analyzed using the KMO and Bartlett methods.

First, the KMO was analyzed: if the KMO<0.6, the data of the scale is not highly reliable. If 0.6 ≤KMO<0.7, the data validity of the scale is moderate. If 0.7≤KMO <0.8, then the data validity of the scale is very good, and if 0.8≤KMO, then the data validity of the scale is very good.

Second, determine P-value: for the analysis of validity, must pass the Butterlett ball test (the corresponding P-value must be <0.05).

Third, In the case of only two questions, the KMO is 0.5 anyway. The validity verification using KMO and Bartlett sphericity tests showed KMO value of 0.879 and KMO value greater than 0.8, showing the study data validity.

3. Principal Component Analysis and Factor Analysis

3.1. Association Analysis of Principal Components with Grey

a) Principal component analysis of the process by which these new variables maintain the original information as much as possible and removes redundant variables, i. e., closely related variables, from all previously proposed variables, and establishes as few new variables as possible so that these new variables are paired and irrelevant [3].

b) Dimensional reduction is the process of retaining the most important features (too many indicators) in the high-dimensional data, removing the noise and the unimportant features, and achieving the purpose of improving the data processing speed. Data dimensionality reduction and its advantages.

I. Making the data set easier to use.

II. Reduce the computational overhead of the algorithm.

III. Remove noise.

IV. Making the results easy to understand.

c) Principal component analysis idea: If there are n samples and P indicators, the size of a sample matrix of n×p can be formed.

$$x = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p) \quad (1)$$

Suppose you want to find A new set of variables, $z_1, z_2, \cdots, z_m (m \leq p)$, and they meet the following requirements.

$$z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \quad (2)$$

$$z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \quad (3)$$

$$z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \quad (4)$$

The data were normalized in order to eliminate the effect of dimension and order of magnitude.

Find the correlation coefficient matrix of index data: Through SPSS software, the correlation coefficient of index data is shown in the following table.

PCA is designed to explain the internal structure between multiple variables through several principal components. Based on the above four AI learning impact evaluation indexes, SPSS factor analysis and principal component analysis are used to obtain the contribution rate of each principal component, rotating component matrix and eigenvector matrix.

Table 1. Coefficient of correlation

	KX1	KX2	KX3	KX4	KX5
KX1	1.000	0.978	0.236	0.603	-0.054
KX2	0.978	1.000	0.368	0.491	-0.214
KX3	0.256	0.357	1.000	0.184	-0.224
KX4	0.601	0.492	0.492	1.000	0.666
KX5	-0.055	-0.214	-0.224	0.666	1.000

Table 2. Total variance was calculated

	Initial eigenvalue			Extract the sum of squared loads		
ingredients	total	Percentage variance	Cumulative (%)	total	Percentage variance	Cumulative (%)
1	3.435	57.251	57.251	73.435	57.251	57.241
2	1.654	27.574	84.824	1.654	27.574	84.724
3	0.805	13.410	98.234			
4	0.064	1.063	99.295			

As can be seen from the table, two principal components were selected in the principal component analysis, with the first contributing 57.214% and the second 27.303%. The cumulative contribution rate was 84.724%, which means that these two principal components cover more than 84.724% of the data and information of the four AI learning impact assessment indicators.

3.2. Based on Clustering Method and Statistical Analysis Model: Factor Analysis-clustering Model

Hierarchical clustering is a hierarchical-based clustering algorithm. There are two methods: bottom-up merging and top-down splitting, and it can be visualized, intuitive and clear.

Silhouette score, also known as the silhouette coefficient, calculates the average distance a from sample i to other samples in the same cluster. The smaller a is, the more likely sample i will be clustered into this cluster. Let the intra-cluster similarity of sample i be a; by calculating the average distance b of all samples from sample i to another cluster C, it is called the similarity between sample i and cluster C. The definition formula is as follows, the closer the score is to 1, the more reasonable the clustering is.

$$s(i) = \frac{b(i) - a(i)}{\max\{\mathrm{a(i)}, \mathrm{b(i)}\}} \quad (5)$$

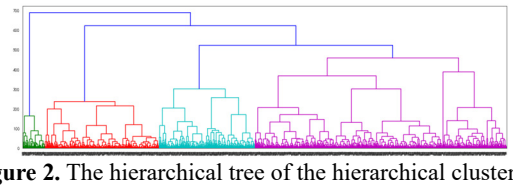


Figure 2. The hierarchical tree of the hierarchical clustering

Parameter discrimination based on the machine learning method, an eight-attribute classification model. A concrete implementation of the gradient lift tree framework (Gradient Boost Decision Tree Framework) was proposed by Chen et al. The basic form is

$$\begin{cases} L(\Phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ \Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2 \end{cases} \quad (6)$$

The loss function during the iterative process is expressed as follows

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}^{(t-1)} + f_t(x_i) + \Omega(f_t)) \cong \sum_{i=1}^{n} \left[l(y_i, \hat{y}^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)) \right] + \Omega(f_t) \quad (7)$$

Different from the ordinary GBDT, XGBoost adds the regularization terms related to the decision tree depth T and the classifier score w after L, which makes the model with superior accuracy more robust. In addition, in a sense, it also has the effect of dimension reduction, and has a better effect on the data of category imbalance.

The above indicators are abstracted, and the factor analysis method is used to further reduce the dimension to find the hidden variables that can describe these eight indicators. The factor load matrix of these eight indicators is plotted as the heat map shown in the figure below

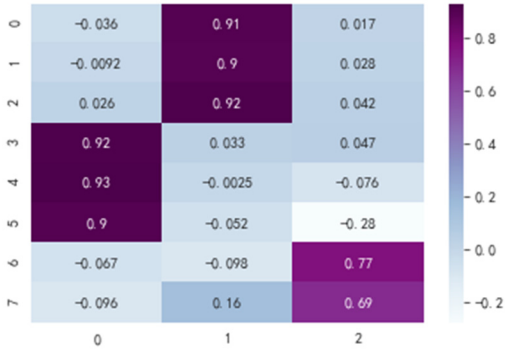


Figure 3. Thermodynamic chart

After calculation, the cumulative variance contribution rate of 0.95 can be obtained when the number of factors is 3. The vertical axis represents each factor from 0 to 7. Factor 0, abstract factor: "Internet mode"; factor 1 abstract factor: "Internet time"; and factor 2, representing "Internet purpose".

After dimension reduction, we plot the cluster map in the three-dimensional coordinate system and plot their projections on the factor 0 and 1 constituent plane in the two-dimensional plane to construct our factor analysis-cluster model.

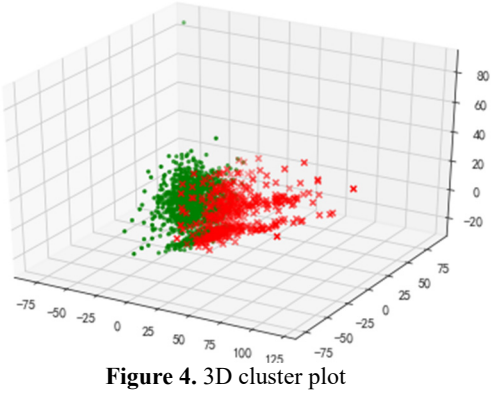


Figure 4. 3D cluster plot

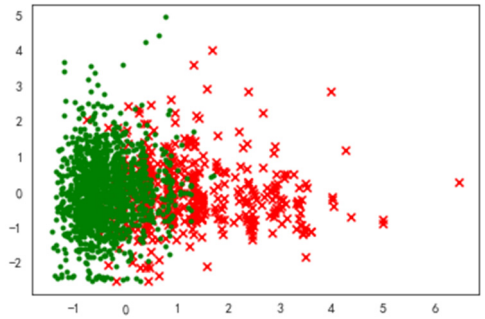


Figure 5. The 2-dimensional cluster graph projection

3.3. Typical Phase Relationship Analysis

Typical correlation analysis is to study how strong the correlation is between two groups (one in each group). In the canonical correlation analysis, a series of linear combinations are selected under certain conditions to reflect the linear relationship between the two groups of variables, and these selected linear combinations are called canonical variables. Both the canonical correlation analysis and the principal component analysis belong to the category of linear analysis.

Suppose that the final impact of AI on learning is U_i, and each influencing factor is V_i. The representative comprehensive variables of both groups are U_i and V_i, and each comprehensive variable is a linear combination of the original variables, namely

$$cov(U_1, U_2) = cov(V_1, V_2) = 0 \quad (8)$$

Now the different indicators are quantitatively analyzed, so as to judge the relationship between them. The correlation coefficient matrix is expressed as follows

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} \quad (9)$$

We selected the linear correlation coefficient

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x})(x_{ki} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x})^2 (x_{ki} - \bar{x}_j)^2}} \quad (10)$$

Table 3. Eigenvector matrix

Index Feature vectors	1	2	3	4	5	6
A	-0.451	-0.007	0.462	-0.447	0.072	-0.615
B	-0.392	0.210	0.510	0.717	-0.092	0.136
C	-0.310	0.557	-0.593	0.113	-0.332	-0.346
D	-0.485	-0.061	-0.329	0.058	0.790	0.160
E	-0.501	-0.052	-0.005	-0.428	-0.410	0.628
F	0.245	0.800	0.258	-0.294	0.290	0.253

Correlation analysis principle: r is between [-1,1], the closer |r| is to 1, the greater the degree of linear correlation of variables; otherwise, the smaller r is. The positive and negative of indicates the positive and negative correlation between variables, which means that one variable increases with the increase of the other variable as a positive correlation, and one variable decreases with the increase of the other variable as a negative correlation. In particular, if r=0, the variables are linearly independent.

Use of the contribution rate formula, $\text{Contribution rate} = \lambda_i / \sum_{i=0}^{p} \lambda_i$

Table 4. A Schematic chart of the contribution rate of each component

Principal component	1	2	3	4	5	6
Contribution rate	0.571	0.179	0.145	0.063	0.024	0.018

Then the feature vector matrix and the contribution rate of each component. Further analysis of Table 3 and Table 4 above shows that the contribution rate of the first three principal components is 89.5%, so they are the main components affecting learning.

$$y_1 = -0.415x_1 - 0.392x_2 - 0.310x_3 - 0.485x_4 - 0.501x_5 + 0.245x_6 \quad (11)$$

$$y_2 = -0.007x_1 + 0.210 + 0.557x_3 - 0.061x_4 - 0.052x_5 + 0.800x_6 \quad (12)$$

$$y_3 = 0.462x_1 + 0.510x_2 - 0.593x_3 - 0.329x_4 - 0.005x_5 + 0.258x_6 \quad (13)$$

They are Internet time, Internet purpose, Internet way.

4. The Impact of Artificial Intelligence on College Students

4.1. Hierarchical Analysis Method (AHP)

Calculating and comparing various factors, we get the weight of different influencing factors. The score was calculated based on the weight matrix, the score was then calculated based on the weight matrix and model corrected using the entropy weight method to obtain the final correction score [4].

Table 5. Summary table of scores

Help papers	Help quizzes	Access information	Chat and make friends	Shop online
0.4399	0.2783	0.4321	0.5633	0.3227

Chatting and dating and being used to help papers score the highest.

4.2. Establishment of the Fischer Discriminating Model

When predicting the impact of artificial intelligence on college students, we mainly consider six questionnaire data: major, grade, whether to use artificial intelligence learning tools, whether to help with homework, whether to help with tests, and whether to help with papers. Data analysis was carried out with the help of SPSS software to determine the proportion of each indicator. Get the discriminant coefficient of Fisher's discriminant function [5].

$$\omega = S_\omega^{-1}(M_1 - M_2 - M_3) \quad (14)$$

Where, S_ω is the within-class total dispersion matrix.

a) Using the spss tool, the mean and standard deviation of the six indicators were calculated, and the difference of the six indicators in each group was analyzed, as shown in the table below

Table 6. Test of difference in class mean

Index	λ	F value	Significance probability
Specialized H1	0.589	20.214	0.003
Grade H2	0.918	2.721	0.072
Use artificial intelligence H3	0.263	22.041	0.089
Help with assignments H4	0.832	5.683	0.229
Help quizzes H5	0.921	1.501	0.051
Help papers H6	0.943	0.506	0.121

b) Difference tests were performed, F is opposite to the relationship, proving that the average difference is obvious. From the table, it can be found that the probability of intra-class mean significance for grade (H2), artificial intelligence (H3), help assignment, help (H4), help test (H5) and help paper (H6) were 0.072,0,0.0289,0.229,0.051 and 0.121, respectively, all greater than 0.05. Only for the professional (H1), the probability of intra-class mean significance does not exceed 0.05, so there are good reasons for the discriminant analysis. Moreover, because the six indices were selected from all the indicators, the linear associations between the six indices were not significant.

Box tests with equal covariance showed that M is 19845, degrees of freedom 1 is 14.321, degrees of freedom 2 is 1123.43, and the probability of significance is 0.000. It is known that the condition that the covariance matrix is equal is that the M value is much greater than 0.05, indicating that all kinds of covariance matrices are equal at the significance level.

c) The discriminant function was analyzed and the function eigenvalue and λ values were calculated. The determination coefficient and constant are shown in the table below.

Table 7. Judgment coefficient and constant quantity

Index	Classify			
	1	2	3	4
Specialized H1	0.75	1.116	0.949	1.009
Grade H2	0.574	0.652	0.549	0.823
Use artificial intelligence H3	4.569	4.225	4.278	4.204
Help with assignments H4	6.848	5.885	5.615	7.631
Help quizzes H5	1.644	1.601	1.435	1.642
Help papers H6	-0.826	0.827	0.345	0.567
Constant	-35.67	-35.626	-39.529	-40.012

The coefficients given by the table above can establish a Fisher linear discriminant model.

$$F_1 = 0.75H_1 + 0.574H_2 + 4.569H_3 + 6.848H_4 + 1.644H_5 - 0.826H_6 - 35.67 \quad (15)$$

$$F_2 = 1.116H_1 + 0.652H_2 + 4.225H_3 + 5.885H_4 + 1.601H_5 - 0.8267 - 35.626 \quad (16)$$

$$F_3 = 0.949H_1 + 0.5749H_2 + 4.278H_3 + 5.615H_4 + 1.435H_5 + 0.345H_6 - 39.529 \quad (17)$$

$$F_4 = 1.009H_1 + 0.823H_2 + 4.204H_3 + 7.631H_4 + 1.642H_5 + 0.567H_6 - 40.012 \quad (18)$$

The indicator values are brought into the four categories as the corresponding variable values. If F 1 is the largest, the student should be classified into category 1. Similarly, again, if F 2 is the largest, the student should be classified as category 2; if F4 is the largest, the student should be classified into category 4. By discrimination, the 2012 samples belonging to category 1 still had 1987 samples belonging to category 1. There were 1003 samples in category 2; 687 samples in category 3 and 920 samples into category 4. Of the eight samples to be identified, three were classified as category 1, three as category 2, and one into category 3. 2 were placed under category 2. Using the Fisher discrimination model established in this paper, the discrimination accuracy rate is 99.82%, which has high credibility, so it can be used as the basis for the discrimination classification of the influence of artificial intelligence on college students.

5. Model Evaluation

Regression models of six factors and learning effects were constructed using multiple regression analysis. After constructing the initial model, the interaction between the two factors was removed and the model was modified. Following stepwise regression and regression analysis of various factors were performed.

a) From the Discrimination (differentiation) or the prediction accuracy: the initial model was modified after the construction of the model; the scientific prediction method of gradual regression analysis was selected, and the regression equation of all factors was introduced by all the effects of the factors on the explanatory variables, resulting in the good prediction accuracy of the model.

b) From the evaluation of the model: =0.0875, it shows that the model is poorly fit. However, because the regression analysis focuses on the significance of the model, the model is still relatively good.

c) From the degree of difficulty in model operation: regression analysis method is simple and convenient to deal with multi-factor model.

6. Conclusion

Through the above analysis and in the future to the development of artificial intelligence, and the understanding of artificial intelligence, cognition and judgment, summarizes the influence of artificial intelligence in the direction of college students' analysis report, respectively expounds the positive and negative effects, as well as to the development of artificial intelligence in the future, application and view.

References

[1] Yonghong Xue. AlphaGo Evolutionary Path and Philosophical Insights[J]. Data,2022(02):14-16.

[2] The usage of the Cronbach Coefficient alpha in the Analysis of the Written Instruments[J]. Revista Médica Electrónica, 2012, 34 (1).

[3] XingBing Bu, Zili Fang,Qiang Feng,Chong Liao,Huizhen Wang. A study on the comprehensive evaluation of air quality based on principal component analysis: an example of 21 cities in Sichuan Province[J].Sichuan Environment,2023,42(03):51-56.DOI:10.14034/j.cnki.schj.2023.03.009.

[4] Xi Chen, Zheng Chen. Pre-assessment of risk of underground cavities in Shanghai roads based on AHP-entropy power method [J]. China Municipal Engineering, 2023(03):102-105+ 109+ 121-122.

[5] Chong Zhong, Qiang Guo.Fisher's discriminant method and its application [J]. Journal of Southwest Jiaotong University, 2008 (01): 136-141.