The Construction of Prediction Model based on Decision Tree-Neural Network Algorithm for Identifying Poverty-Stricken College Students

Yuncong Zeng[1, *], Yifan Han[2]

[1]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu City, Sichuan Province, China
[2]School of Public Affairs and Administration, University of Electronic Science and Technology of China, Chengdu City, Sichuan Province, China
[*]Corresponding author
life_zengyc@uestc.edu.cn, 13051523565@163.com

Abstract

In the context of "three comprehensive education" in campus, the support for economically disadvantaged students in campus is still one of the key tasks that need to be solved. With the limited resources of financial support for poverty-stricken college students, the identification of these students faces serious challenges. In this essay, a prediction model based on decision-tree and neural network algorithm is designed to predict and analyze students' poverty recognition. After testing and validation, it proves the effectiveness of the model applied to the prediction of college poverty-stricken college students. Based on the author's database, the recognition accuracy of the model can reach 97.52% at present, and the prediction results of student poverty recognition can be realized with data update.

Keywords: *Decision Tree-Neural Network, Poverty-Stricken College Students, Prediction*

1 INTRODUCTION

As an important part of the construction of China's financial aid system for poverty-stricken college students, the national financial aid system should not only give full play to the important role of national economic policies, but also give full play to the role of financial aid work in the "three comprehensive education". The large size of poverty-stricken college students in campus has caused a series of family, school and social problems due to the overlapping of economic, psychological and developmental problems. Given the limited resources for financial support for poverty-stricken college students, the identification of poverty-stricken college students faces serious challenges. Hegedus (2018) [5] mentions that educators who serve students living in poverty sometimes face serious educational and economic barriers. Exploring a scientific, economical and operable method for identifying poverty-stricken college students has become the first issue that needs to be addressed in the work with poverty-stricken college students.

To address the above issues, this paper firstly identifies and classifies the sponsored students based on the author's student work experience and relevant research and seminar analysis, and then obtains the students' emotional feature vectors by identifying the text-based databases such as the basic student database and the interview database, and combines the digital data sets such as the students' poverty certificate database, the interview database, the basic student information database and the school life database. The student feature vector is obtained by stitching together with the digital data sets such as the student poverty certificate database, the student talk database, the student basic information database and the school life database. Using the student feature vector as input, a prediction model based on decision-tree and neural network algorithm is designed to predict and analyze the poverty recognition of students.

2 DATA DESCRIPTIONS

2.1 *Definition and Classification*

At present, there is no clear definition of poverty-stricken college students in campus, so by referring to relevant literature and combining with the author's

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 1287-1293, 2023.
https://doi.org/10.2991/978-94-6463-040-4_191

working experience, the following classification of poverty-stricken college students is made in this essay.

As shown in Table 1 below, it is the comprehensive three-dimensional evaluation and identification index system for poverty-stricken students in campus established in this paper based on relevant research. The types are divided into economic situation, family situation, personal situation and other situations.

Table 1: Comprehensive three-dimensional evaluation and recognition index system for poverty-stricken college students in campus

Type	Positive Terms	Negative Terms
Economic Situation	Normal, rich, No debt, economically developed area	Rural, farming, low annual income per capital
Family Situation	Average economic family, high level of education, parents have proper jobs	Single parents, divorce, domestic violence, parental disability, parental delinquency
Individual situation	Good health, stable emotion, in good academic and living condition and perform well in school.	Orphans, minorities, disabilities, body part information, diabetes, heart disease, dwarfism, malnutrition, somatization reactions, etc
Other conditions	No disasters or accidents in daily life	Poverty caused by chance, sudden floods, droughts, poverty caused by irresistible factors, etc

The table below shows the comprehensive three-dimensional evaluation index system for identifying poverty-stricken college students in campus, which is divided into four categories: economic situation, family situation, personal situation and other situations. Sociodemographic and political factors should not be ingored [2] and someone analyzes that students academic performances will be influenced by insufficient socioeconomic requirements [6].

Table 2 is a supplement to the conditions for identifying poverty-stricken college students in campus that contains special situations of family or individual. According to various cases, this table will analyze and score each student's specific family situation to improve the relevant data information of the comprehensive three-dimensional evaluation and identification index system for poverty-stricken college students in campus in Table 1.

Table 2: Supplement to the conditions for identifying poverty-stricken college students in higher education

Conditions	Visible Observations	Requirements and instructions
Mandatory conditions (70%)	Orphan (A1)	Both parents dead, no guardian
	Children of martyrs (A2)	Single parents, domestic violence

	Core Member Major Illness(A3)	reference to medical history years of experience, severity of illness, treatment costs
	Single Parent Families(A4)	One of the parents is dead or missing
	Divorced Families (A5)	Parents are divorced
	Education of family members (A6)	elementary school students in remote areas
	Disaster Area（A7）	Areas where major natural disasters have occurred
	laid off, low income (A8)	Valid for immediate family members
	Old, young, poor and marginalized students (A9)	Old revolutionary areas
	Incidental Events（A10）	Traffic accidents, fires, etc
	Other Situations（A11）	Disabled or mentally handicapped
Observation conditions (30%)	Life Consumption (Campus Stored Value Card)	This section should have a negative scoring scheme
	Communication or information costs (B1)	the higher the applicant's cost, the lower the score
	Recreation or socializing (B2)	

2.2 Description of Student Database

Commonly used methods for identifying poverty-stricken college students: poverty certificate identification method, family household and place of origin analysis identification method, counselor's intuitive judgment identification method, democratic evaluation identification method, school consumption level identification method, resident minimum subsistence line comparison identification method, family per capital income analysis identification method, comprehensive analysis identification method. Based on the author's experience in student work and the analysis of relevant research seminars, the following database methods are to be used in this study.

(1) Poverty certificate database. One category is the students whose families are mainly selected inside the village through precise poverty alleviation, the degree of poverty is recorded, the poverty file is established, and the families who get the poverty card. Free or reduced-priced lunch is shown as a proxy measure for identifing student poverty [3]. There is also a category of low income student families, which are families whose income level is lower than the city's monthly average and then have state subsidies. The poverty certificate database is the most effective and powerful proof to judge poverty-stricken college students.

(2) Talking database. Decree 43 of the Ministry of Education clearly puts forward that college counselors should strive to become life tutors for students to grow and become successful and heartfelt friends for healthy life, and heart-to-heart talk becomes an important means for counselors to timely and effectively understand the ideological and political status and study and life status of each student, such as the author's unit clearly requires counselors to submit no less than 400 records of student heart-to-heart talk in the system every semester. On the one hand, you can grasp the students' family economic situation and occasional events through the talk database,

on the other hand, you can grasp the students' school dynamics, life consumption level, etc. through the talk. Simultaneously, according to the conversation of dormitory students and class committee leaders, we can get a side view of each student's family economic situation and living consumption level. Therefore, the database of heart-to-heart talks can be used to identify and track students in difficulty.

(3) Students' basic information database. Student information is collected and updated through online questionnaires and system filling before students enter school or at the beginning of each semester, mainly including students' home address, health of family members, statistics of disaster losses suffered in the past three years, per capita annual family income, occasional events and other related contents, all of which can be used as data.

(4) On-campus life database. The relevant data of students on campus can be used as reference database. Students' one-card consumption data, some campus have realized the identification of poverty-stricken college students based on students' one-card consumption data; students' punch card record data of entering and leaving the school, students who enter and leave the school regularly may have abnormalities; students' awards and grants won, discipline competitions won, and work-study during their school years. Accordingly, the poverty of a campus is related to many of the student outcomes [1]

(5) Special student database. Through various means, counselors and other thinking staff clearly classify some special students, i.e., grasp their characteristic labels (such labels are rare and unevenly distributed in the sample).

Based on the above data set, a big data sink of students is formed and used to identify poverty-stricken college students. Howard (2001) [4] concludes that helping poverty students in college is only half the battle. Similarly, categorizing student poverty factors is a way to better target various students in poverty, and then find efficient way to reduce the impact of poverty issues on students in the future.

3 STUDENTS IDENTIFICATION MODEL

3.1 *Student Data Processing*

3.1.1 Downscaling and Visualization of Student Data

In this essay, the t-SNE algorithm is used to downscale and visualize the student data so that the point probabilities corresponding to the high-dimensional space and the low-dimensional space are the same. t-SNE is essentially an embedding model that can map the data in the high-dimensional space to the low-dimensional space and retain the local characteristics of the data set. Form 2 has 13 entries, which are processed as 13 dimensions and reduced to 4 dimensions using the t-SNE algorithm, consistent with Form 1.

3.1.2 Pre-Processing of Student Data

In this essay, the sentiment analysis of the database text information is used to obtain the correlation degree matrix of students and special students, and the data with the highest relevance is used as clusters.

Firstly, the semantic data of students' conversations are read, and all the data are divided into words, and the obtained division results are intersected with the sentiment dictionary to get a new sentiment word dictionary. As shown in Table 1, for each student's evaluation, find the emotion words in their evaluation to record their positive and negative and record the category they belong to, add 1 to the number of positive words if positive words exist, add 1 to the number of negative words if negative words are contained, and multiply the value of the emotion word by different coefficients according to the type of different adverbs if degree adverbs exist before the emotion word. The weights of degree adverbs: slightly, generally, very, extremely, and extremely, are assigned 0.8, 1, 1.2, 1.4, and 1.6, respectively, to obtain the final vector. The assignment methods in Table 2 are: A1, A2, A4, and A5 add 5 when the condition is satisfied, and 0 when it is not; A3, A6, A7, and A8 are 6 levels, corresponding to the corresponding level from 0-5 points (rounded); A10, A11, B1, and B2 are assigned with consecutive values in the range of 0-5 points. Then, the formula for calculating each student in the 4-dimensional poverty assessment is as follows:

$$Final_{score} = \alpha_1 V_1 + \alpha_2 V_2 \qquad (1)$$

In this formula, V_1 denotes the vector dimension of Table 1, and $\alpha_1$is the corresponding weight coefficient of Table 1. V_2 denotes the vector dimension of Table 2, and α_2 is the corresponding weight coefficient of Table 2.

Take a freshman undergraduate student in the author's unit as an example, his family is normal, but comes from an economically underdeveloped area with a low student saving card balance and relatively serious family debt. The family has an average level of education. The father has a disability with a normal job, and the mother is retired. She is unhealthy with high blood pressure and high blood sugar. Because this family encounters a severe earthquake, it results that $V_1 = [-2, -0.8, -2.4, -1.6]$. After completing all questions in Table 2 and then using the t-SNE algorithm to reduce the dimensionality. Finally, The data results in Table 2 shows that $V_2 = [-1.9, -0.4, -2.5, -1.5]$. Finally, the 4-dimensional results of this student in poverty assessment is $Final_{score} = [-1.98, -0.72, -2.42, -1.58]$.

To facilitate model training, of the economically disadvantaged student pool is used as training set, 10%

as the validation set, and the remaining 10% as the test set. In order to solve the problem that the difference between the number of normal and abnormal samples is too large, this paper performs data augmentation on different abnormal data so that the number of abnormal and normal students in different categories is approximately equal to improve the accuracy of the model training results.

3.2 Integrated Learning

3.2.1 Kmeans Clustering

Firstly, the student feature vector is clustered using Kmeans clustering and the number of clusters is set to 3. 20 clusters are performed and the results of each cluster are saved. The clustering effect is shown in Figure 1. For ease of demonstration, only a part of the student feature vectors is used in Figure 1.

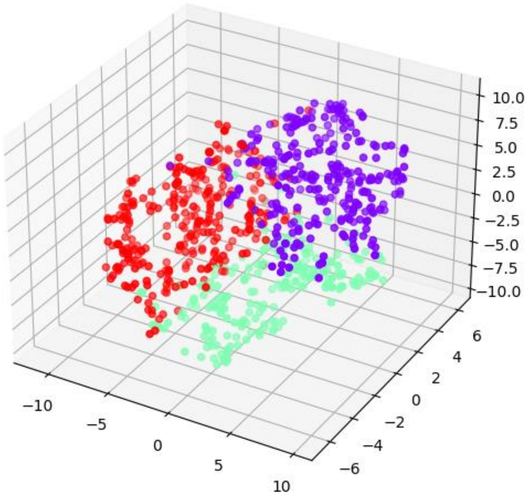

Figure 1: Schematic 3D effect of feature vector clustering for poverty-stricken college students

In this paper, the students who applied for the database of identification of students with financial difficulties were finally identified as special difficulties, difficulties and general difficulties. Therefore, the number of clusters is set as 3, and the data reaches steady state after 20 clusters. The final probability of each type of each student in this paper is the mean value of the 20 times clustering results. And the accuracy was calculated on the test set.

3.2.2 Construction of Decision-Tree Algorithm

In this paper, the decision-tree algorithm of machine learning is utilized, and the generation algorithm is selected as ID3, which is a tree structure in which each internal node represents a judgment on an attribute, each branch represents an output of a judgment result, and finally each leaf node represents a classification result. In a given database of poverty-stricken college students identification, each poverty-stricken college students sample has a set of attributes and a classification result, that is, the classification result is known, then a decision-tree is obtained by learning these samples, and this decision-tree is able to give the correct classification for the new data. As shown in Figure 2, the decision-tree algorithm is divided into 3 levels, and the classification levels are Slight, Normal, and Extreme.

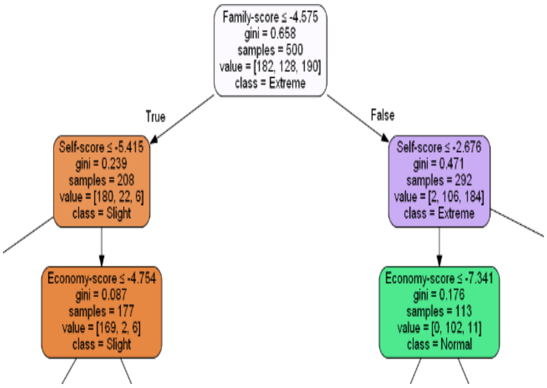

Figure 2: Partial structure of decision-tree algorithm

3.2.3 Construction of Neural Network Learning Algorithm

A 4-layer fully connected BP neural network is used as another input model for integrated learning, in which the feature values of the four dimensions of students are used as input, the three poverty levels of students belonging to different types are used as output, and for data with real labels, the likelihood of the type corresponding to their labels is 1 and the likelihood of the remaining types is 0. The reLU function is used as the activation function. The structure of the neural network algorithm is shown in Figure 3 below. MLP Classifier is a supervised learning algorithm, and the following figure shows the MLP model with only 1 hidden layer , the left side is the input layer and the right side is the output layer. Each neuron in the input layer is fed with four different types of features from the original data, and these features are fed in the form of a matrix. Each neuron in the hidden layer represents the data updated once for the poverty-stricken college students parameters, while each layer has N neurons indicating that the features of the input data are extended to N. Each neuron in each layer can have a different representation. Then the output of the hidden layer is:

$$f(W^{(1)}x + b^{(1)}) \tag{2}$$

In this formula, $W^{(1)}$ is the weight (also called the connection factor), and $b^{(1)}$ is the bias. The function can be the commonly used sigmoid function or tanh function:

$$sigmoid(a) = 1/(1 + e\hat{}(-a)) \tag{3}$$

$$\tanh(a) = (e^{a} - e^{-a})/(e^{a} + e^{-a}) \tag{4}$$

The output of the output layer is $softmax(W^{(2)}x^{(1)} + b^{(2)})$, $x^{(1)}$ indicates the output of the hidden layer: $f(W^{(1)}x + b^{(1)})$. The entire model of the MLP is shown here and the three-layer MLP can be expressed in the following equation:

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{1}x))) \qquad (5)$$

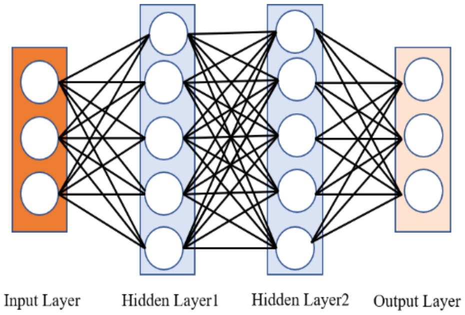

Figure 3: Neural network structure diagram

3.2.4 Analysis of Results

By comparing in decision-tree algorithm, neural network algorithm and neural network algorithm combined with decision tree, the accuracy change of different prediction assessment models in the process of neural network iteration is calculated in the process of 10% data as test set is shown in Figure 4 below. Through this figure, it is obvious that the accuracy rate of the decision tree-neural network integrated learning poverty-stricken students in campus identification model based on the database of this paper can be 97.52%, which is better to achieve the identification of poverty-stricken college students. With the update of the database, the model can effectively complete the work of financial aid education in campus.

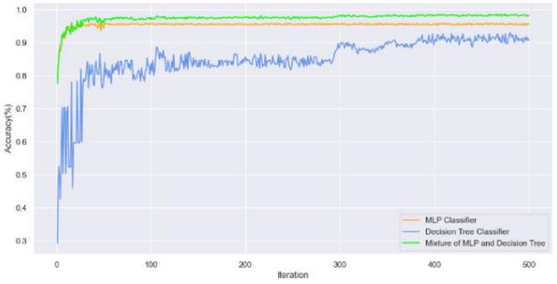

Figure 4: Identification of poverty-stricken college students based on decision tree and neural network integrated learning

4 CONCLUSIONS

Based on the confusing management situation of the identification of students with financial difficulties in campus, and after researching the complexity and inconvenience of the identification of poverty-stricken college students in campus, this study designed and proposed a prediction model for the identification of poverty-stricken college students in campus based on decision tree and neural network algorithm to predict and analyze the identification of poverty-stricken college students. Based on the author's database, the recognition accuracy of the model can reach 97.52%, and the prediction results can be updated with the data. To a certain extent, it can assist college management workers in making certain decisions on poverty recognition.

REFERENCES

[1] Battistich, V., Solomon, D., Kim, D. I., Watson, M., & Schaps, E. (1995). Schools as communities, poverty levels of student populations, and students' attitudes, motives, and performance: A multilevel analysis. *American educational research journal*, 32(3), 627-658.

[2] Griffin, W. E., & Oheneba-Sakyi, Y. (1993). Sociodemographic and political correlates of university students' causal attributions for poverty. Psychological Reports, 73(3_part_1), 795-800.

[3] Greenberg, E. (2018). New Measures of Student Poverty: Replacing Free and Reduced-Price Lunch Status Based on Household Forms with Direct Certification. Urban Institute.

[4] Howard, A. (2001). Students from poverty: Helping them make it through college. About Campus, 6(5), 5-12.

[5] Hegedus, A. (2018). Evaluating the Relationships between Poverty and School Performance. *NWEA Research.* NWEA.

[6] Taha, W. (2022). The influence of poverty on the academic performance of students within the School of Education of Makerere University in Kampala, Uganda (Doctoral dissertation, Makerere University).