Received: 16 January 2021 | Revised: 20 March 2022 | Accepted: 15 September 2022

DOI: 10.1002/asi.24718

RESEARCH ARTICLE

JASIST WILEY

Lost at starting line: Predicting maladaptation of university freshmen based on educational big data

Teng Guo[1] | **Xiaomei Bai**[2] | **Shihao Zhen**[1] | **Shagufta Abid**[1] | **Feng Xia**[3]

[1]School of Software, Dalian University of Technology, Dalian, Liaoning, China

[2]Computing Center, Anshan Normal University, Anshan, Liaoning, China

[3]Institute of Innovation, Science and Sustainability, Federation University Australia, Ballarat, Victoria, Australia

Correspondence
Feng Xia, Institute of Innovation, Science and Sustainability, Federation University Australia, Ballarat, Vic. 3353, Australia.
Email: f.xia@ieee.org

Abstract

The transition from secondary education to higher education could be challenging for most freshmen. For students who fail to adjust to university life smoothly, their status may worsen if the university cannot offer timely and proper guidance. Helping students adapt to university life is a long-term goal for any academic institution. Therefore, understanding the nature of the maladaptation phenomenon and the early prediction of "at-risk" students are crucial tasks that urgently need to be tackled effectively. This article aims to analyze the relevant factors that affect the maladaptation phenomenon and predict this phenomenon in advance. We develop a prediction framework (MAladaptive STudEnt pRediction, MASTER) for the early prediction of students with maladaptation. First, our framework uses the SMOTE (Synthetic Minority Oversampling Technique) algorithm to solve the data label imbalance issue. Moreover, a novel ensemble algorithm, priority forest, is proposed for outputting ranks instead of binary results, which enables us to perform proactive interventions in a prioritized manner where limited education resources are available. Experimental results on real-world education datasets demonstrate that the MASTER framework outperforms other state-of-art methods.

1 | INTRODUCTION

The importance of higher education is self-evident. After finishing higher education, teenagers can better adapt to society by increasing professional knowledge and practical skills (Burrows, 2018; Chen et al., 2017; Kim et al., 2018). However, it turns out to be a stressful experience for freshmen to attend a university (or college) due to a completely unknown environment and autonomous learning style (Auerbach et al., 2018; De Vos et al., 2015; Eisenberg, 1970; Gewin, 2018), which leads to the emergence of maladaptation phenomenon, as shown in Figure 1. The maladaptation phenomenons bring serious negative effects on students' campus life, including academic difficulties (Wintre et al., 2006), emotional problems (Horgan et al., 2016; Puff et al., 2016), and dropping out (Credé & Niehorster, 2012; Gravini Donado et al., 2021). If colleges cannot offer interventions in time, the situation of these students will deteriorate, and some extreme behaviors may occur, such as suicide (Bruffaerts et al., 2019; Graetz et al., 2018; Patton et al., 2018). Understanding the nature of the maladaptation phenomenon and the early prediction of "at-risk" students is the key to solving this problem. In this case, the purpose of this research is to analyze the relevant factors that affect the maladaptation phenomenon and design a framework to predict "at-risk" students in advance.

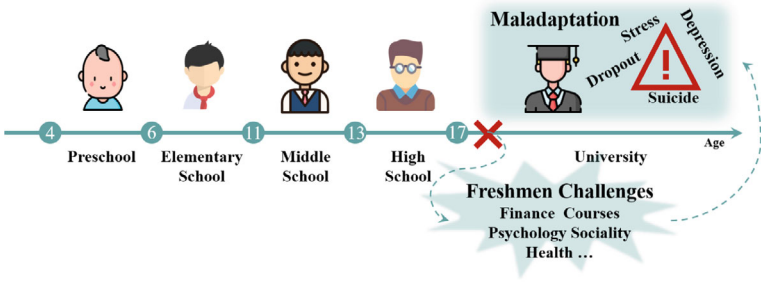

FIGURE 1 The illustration of college freshmen's maladaptation. This figure shows the process of human receiving education from birth to university and the emergence of college freshmen's maladaptation

However, predicting students who may suffer from maladaptation in advance faces tremendous challenges (Gravini Donado et al., 2021; Yao et al., 2013). Studies demonstrate that the status of college freshmen could be impacted by the interaction of biological, psychological, and social factors (Bruffaerts et al., 2018; Liang et al., 2020). Understanding such a complex phenomenon comprehensively is not easy. Moreover, current research in related fields mainly collect data through questionnaires and self-reports which are not only easily influenced by the subjective feelings of participants, but also time- and cost-consuming and hardly applicable to large-scale students (Cao et al., 2018; Johnson & Smith, 2017). In addition, current research related to college freshmen maladaptation is biased toward analyzing correlations between features rather than predictions (Ikui, 2019). In this context, data-driven prediction of college freshmen maladaptation is still an open problem.

Campus life is information-intensive (Bai et al., 2021; T. Guo et al., 2020), and thanks to the popularization of education management systems, massive amounts of education-related data are recorded and stored (T. Guo et al., 2021; Zhang et al., 2019). Meanwhile, data mining technology for massive data has also achieved rapid development (Gupta & Chandra, 2020). Both of these bring us an unprecedented opportunity to solve problems related to the maladaptation of college freshmen. However, new challenges await researchers who concern about this issue. First, the maladaptation phenomenon is not easy to be directly quantified. Finding an indirect indicator that is easy to collect at scale to help quantify the maladaptation phenomenon is the key to the solution. Second, algorithms based on deep learning are the mainstream technologies in the field of big data processing, currently. The inexplicability of this technology makes it difficult to gain the trust of all stakeholders and implement it (T. Guo et al., 2021; Xu, 2020). Moreover, the freshmen with the maladaptation phenomenon are always the minority, so this data mining process inevitably suffers data label imbalance (H. Guo et al., 2017). Therefore, an effective framework that is targeted to solve the above problem is urgent to design.

In this article, we are devoted to analyzing the relevant factors that affect the maladaptation phenomenon and predicting students with maladaptation through tackling the challenges mentioned above. First, previous studies show that academic performance is an important indicator for students' adaptation to college life (Beyers & Goossens, 2002; Gravini Donado et al., 2021; Soledad et al., 2012). In this case, the maladaptation phenomenon is measured by abnormal academic performance fluctuation before and after college admission in our research. Second, we profile college freshmen from three perspectives: intelligence, psychology, and demography, and analyze the correlation of these features with the maladaptation phenomenon. Third, we propose a MASTER (MAladaptive STudEnt pRediction) framework to predict students with the maladaptation phenomenon, which includes two critical steps. First, a variant of random forest, named priority forest, is proposed to predict the students with maladaptation phenomenon. We utilize the optimization strategy of learn-to-rank algorithms to assign weights to the base classifier, which improves its prediction performance while preserving the interpretability of the algorithm. Second, we apply the SMOTE (Synthetic Minority Oversampling TEchnique) algorithm to settle data label imbalance, due to that the number of students with maladaptation is much smaller than normal students. The experiment results demonstrate that our algorithm outperforms other popular algorithms remarkably, including random forest, support vector machine (SVM), and deep neural networks (DNN). The whole experimental process is shown in Figure 2.

Our contributions could be generalized as follows:

- We use academic performance fluctuations to automatically identify college freshmen with the maladaptation phenomenon, and try to predict them through data-driven methods.
- We develop a novel ensemble classification algorithm (priority forest). As a variant of random forest, it generates ranks instead of binary results that prioritize the maladapted students possible.

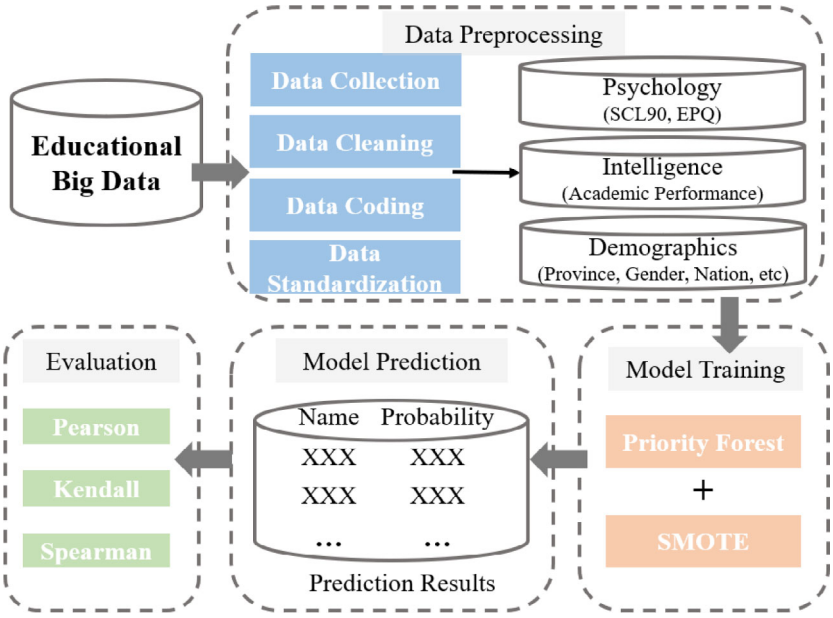

FIGURE 2 The MASTER framework. This figure contains four parts of the experiment: Data processing, model training, model prediction, and evaluation

- We conduct extensive evaluations on real-world education data (intelligence, psychology, and demography) to demonstrate the effectiveness and precision of the priority forest and MASTER framework, respectively.
- We exploit the experimental results to verify the maladaptation phenomenon's predictability, which contributes to educational psychology studying and education policymaking.

This article is organized as follows. In the next section, we review the related work. The problem formulation is presented in Section 3. In Section 4, we introduce the proposed framework in detail. In Section 5, we describe the dataset used in this study. In Section 6, we show the experiment results. In the final section, we present the discussion and conclusion of our work.

2 | RELATED WORK

This section provides an overview of the related research.

2.1 | Maladaptation phenomenon of college students

Existing research related to the maladaptation phenomenon mainly focuses on two groups of students: international students and college freshmen.

First, scholars explore the adaptation problem of international students from multiple aspects. Y. Wang et al. (2018) explored the temporal patterns of students' psychological and socio-cultural adaptation based on 169 international students. Meanwhile, Lee et al. (2018) also focused on international students. They explored the relationships between the seriousness of leisure activities, social support, and school adaptation. Gu and Usinger (2021) focused on international Chinese graduate students in the United States and designed experiments to explore their intercultural adaptation. Panich et al. (2021) explored the maladaptation phenomenon of international students from a cultural perspective and attempted to solve this problem through cultural assimilator. Oh and Butler (2019) focused on the adaptation of international college freshmen in a foreign university from the perspective of information behavior. They identified differences in international college freshmen's use of information sources during the adjustment period. Sahão and Kienen (2021) discussed the relationship between students' mental state and their maladaptation through a literature review.

Second, scholars explore the adaptation problem of college freshmen from various aspects. Deshpande et al. (2009) conducted a study to explore the maladaptation of eating habits of college freshmen. The results proved the correlation between eating habits and physical health. Butler et al. (2004) also performed a study to explore the maladaptation of the diet, physical activities, and bodyweight changes of college freshmen. The results displayed that they need interventions in these aspects. Bruffaerts et al. (2018) performed a study to investigate the extent to which mental health problems are associated with college freshmen's academic maladaptation. The result showed that mental health problems are associated with lower intellectual functioning. Clark and Cundiff (2011) conducted a study to explore the relation between the

grade point averages of first-year college students and retention rates. Krajniak et al. (2018) carried out a study to explore the relationship between personality disorder traits, emotional intelligence, and college adaptation from psychology. The results suggested an alternative model implicating emotional intelligence as a mediator of the relationship between personality difficulties and college adjustment.

Some fatal shortages exist obviously in current research. First, maladaptive students at the beginning of university life should be defined by comparing their performance before and after college admission, which plays a role like control groups, rather than making a decision only according to the performance at the beginning of college life. Second, most known studies in this domain are mainly based on questionnaires with sample sizes usually scaling from dozens to hundreds. In addition, experimental bias exists in these studies since subjects would like to report desirable personal information instead of disapproved behaviors. Moreover, predicting the maladaptation phenomenon in advance is a key step in solving the problem, but it is rarely mentioned in the current related research.

2.2 | Variants of random forest

Random forest proposed by Breiman (2001) is an ensemble algorithm that combines tree predictors and uses ensemble strategies to improve classification accuracy. In the past few years, this algorithm has been widely used in various scenarios and achieved good performance (Resende & Drummond, 2018; Shaik & Srinivasan, 2019). Many studies have been carried out to improve its performance (Cui et al., 2015; Gieseke & Igel, 2018; Sathe & Aggarwal, 2017). Meta random forest is proposed by Bonissone et al. (2010) by combining bagging and boosting approaches to improve the performance of random forest. They utilized the random forest as the base classifier with the bagging approach and the boosting approach, separately. Rodriguez et al. (2006) proposed an algorithm named rotation forest that generates classifiers based on feature extraction. They preserved the variability information in the data through principal component analysis to improve performance. Bernard et al. (2012) improved random forest based on an adaptive tree induction procedure. They proposed a dynamic random forest that guides the tree induction so that each tree involved will complement as much as possible existing trees in the ensemble. Speiser et al. (2019) proposed a random forest method to combine clustered results and longitudinal outcomes called binary mixed model (BiMM) forest. The accuracy of BiMM forest is higher than the existing methods for predicting the outcome of new subjects. Zhou and Qiu (2018) proposed a random forest-based label ranking method. They developed a novel two-step rank aggregation strategy to effectively aggregate neighboring rankings discovered by the random forest into a final predicted ranking. The main goal of the algorithm mentioned above is to make each learner in the forest more differentiated (e.g., bagging series algorithm) or to make the newly trained learner more robust (e.g., Adaboost series algorithm) by using the weight of data sets, and then to integrate all learners using various set strategies to get the final strong learner.

3 | PROBLEM FORMULATION

In this section, we introduce notations involved in this article and then formally define the problem in this work. In a university, let $\boldsymbol{m} = \{1, 2, ..., m\}$ denotes the set of majors (i.e., programs). The set of students in every major is defined as $\mathcal{N} = \{\mathcal{N}_1, \mathcal{N}_2, ..., \mathcal{N}_m\}$. As mentioned before, maladaptation phenomena can hardly be measured explicitly, so we use the abnormal fluctuation of academic performances as a proxy. The details are shown as follows.

Maladaptation definition: Based on the assumption that students with maladaptation tend to perform a dramatic decline in academic performance at the beginning of college life, we use abnormal fluctuation of academic performances to recognize students' adaptation conditions. Note that, all students' academic performance is represented by their rankings. For example, if Jack's academic performance exceeds 98% of the students, his academic performance is 0.98. For students in major m, they are first ranked by their college entrance exam scores, denoted as R_{before}. Then, they are ranked by their performance of the first semester exam in college (grade-point average [GPA]), denoted as R_{after}. The academic performance fluctuation of student i, denoted by α_i, is defined as follows:

$$\alpha_i = R_{\text{after}-i} - R_{\text{before}-i}, \tag{1}$$

α_t is a threshold value defined by a stochastic model, which will be described in Section 6.1. $y_i \in \{0, 1\}$ represented the maladaptation condition is defined as follow:

$$y_i = \begin{cases} 0 & \alpha_i < \alpha_t \\ 1 & \alpha_i \geq \alpha_t \end{cases}. \tag{2}$$

Prediction problem: For student i, we define three types of features, including intelligence features (college entrance exam performance), psychology features

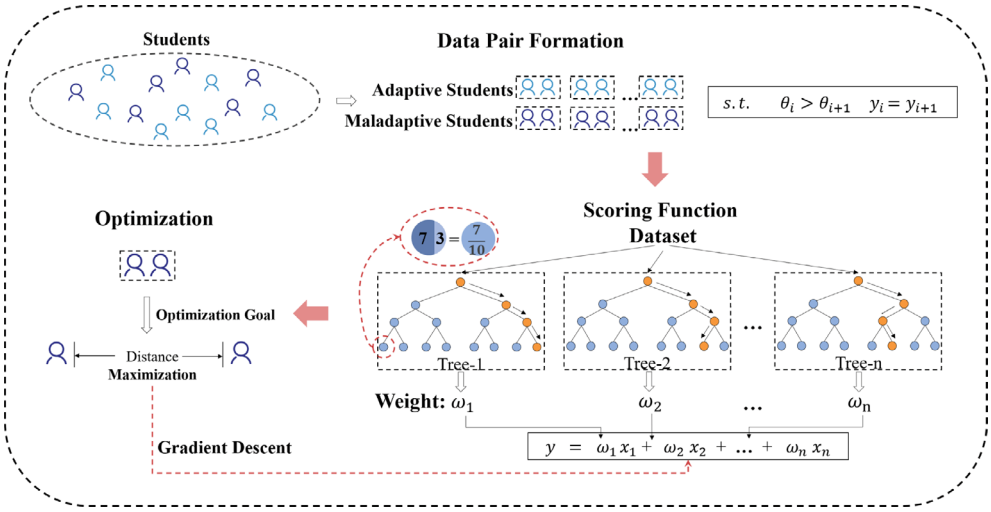

FIGURE 3 The illustration of the priority forest algorithm. We illustrate its three steps, including data pair formation, scoring function, and optimization

(psychological test) and demography features. First, we use the total score of the college entrance exam to represent students' intelligence feature, denoted as $a \in \mathbb{R}$. Second, we use the results of psychological tests at the beginning of college life to represent students' psychological features, denoted as $\mathbf{p} \in \mathbb{R}^n$, where n represents the number of psychological test scales. Moreover, for the demography features, we further propose four categories of features based on raw demographic information, denoted by $\mathbf{d} \in \mathbb{R}^4$, where 4 is represented four features extracted from the demographic data (details are introduced in Section 6.2.2).

Given the feature vector $\boldsymbol{x}_i = [a_i, \boldsymbol{p}_i, \boldsymbol{d}_i]$, we predict the probability that student i suffers from maladaptation.

4 | METHODS

In this section, the MASTER framework, including the priority forest algorithm and SMOTE algorithm (shown in Figure 2), is described in detail.

4.1 | Prediction model

In reality, due to the huge disparity between the number of students and the number of teachers, the severity of the maladaptation can provide a priority order for the work of relevant teachers. For example, if a teacher is responsible for 30 students, and the algorithm tells he\her that 10 of them will feel maladaptive to college life. Since he\she did not know the severity of the maladaptive students, he\she could not know which student needed to be the first to receive the psychological intervention. Inspired by such a special situation, we propose a variation on the random forest, named priority forest, with the aid of optimization ideas of the learning-to-rank algorithm to assign weights to base classifiers by highlighting their interindividual differences. The priority forest algorithm includes three steps: data pairs formation, scoring function design, and optimization, shown in Figure 3, and details will be introduced as follows.

4.1.1 | Data pairs formation

To take into account the priority of students in the optimization process, we first build the set of data pairs based on the original dataset as follows:

$$\mathcal{P} = \left\{\mathcal{P} | \mathcal{P} = (\mathcal{N}_i, \mathcal{N}_{i+1})\ \mathcal{N} \in \mathcal{N}\ ,\ \theta_i > \theta_{i+1}\ ,\ y_i = y_{i+1}\right\},$$

where θ_i represents the similarity between feature $\mathbf{x_i}$ and label y_i and is defined as $\theta = f(\alpha_i)$. In this research, f is an exponential function.

4.1.2 | Scoring function design

Given a characteristic feature vector $\mathbf{x} \in \mathbb{R}^p$, a scoring function $f : \mathbb{R}^p \to \mathbb{R}$ can map the label of samples to a score s.

In this part, a random forest algorithm is used to give a score. We transfer the binary results of each base classifier to continuous results utilizing the probability of each category. For example, there are N samples in a leaf node in a binary classification, and n of them are labeled as positive. The probability of this leaf node to predict a positive case is defined as:

$$h_m(x_j) = \frac{n}{N}. \tag{3}$$

Then, the output $S(x_j)$ is defined as follows:

$$S(x_j) = \sum_{m=1}^{T} \omega_m h_m(x_j), \tag{4}$$

where $h_m(x_j)$ is the classification result of each base classifier and ω_m represents each base classifier's weight and is randomly obtained at initialization. m is the number of base classifiers. T is the number of base classifiers.

4.1.3 | Optimization

In this step, we design an optimization strategy to optimize the weights of classifiers ω_m. Inspired by the loss function of the learning-to-rank algorithm, we create our loss function to distinguish each pair based on the initial idea that maximizes the scores difference of two students ranked next to each other (shown in Figure 3). A synchronous factor $\Omega_i = \Delta\theta_i$ is proposed to standardize each pair's weight.

Finally, we propose such loss functions in this study:

$$\arg\min_{\omega} E = \frac{1}{N}\sum_{i=1}^{N} E_i, \tag{5}$$

$$E_i = \frac{1}{2}\left(1 - M\left(\frac{s_i - s_{i+1}}{\theta_i - \theta_{i+1}}\right)\right)^2, \tag{6}$$

$$M(x) = \frac{1}{1+e^{-x}}.$$

Next, we update the weight in the target's negative gradient direction based on the gradient descent strategy. We set the learning rate η, which controls the updated step in each iteration of the algorithm:

$$\Delta\omega_m = -\eta\frac{\partial E_i}{\partial \omega_m}, \tag{7}$$

$$\frac{\partial E_i}{\partial \Delta y_i} = M\left(\frac{\Delta y_i}{\Omega_i}\right)\cdot\left(1 - M\left(\frac{\Delta y_i}{\Omega_i}\right) + M\left(\frac{\Delta y_i}{\Omega_i}\right)^2\right).$$

The complete updated equation is obtained:

$$\Delta\omega_m = \begin{array}{l} -\eta\cdot(h_m(x_i) - h_m(x_{i+1}))\cdot M\left(\frac{\Delta y_i}{\Omega_i}\right) \\ \cdot\left(1 - M\left(\frac{\Delta y_i}{\Omega_i}\right) + M\left(\frac{\Delta y_i}{\Omega_i}\right)^2\right). \end{array} \tag{8}$$

The whole process is shown in Algorithm 1.

Algorithm 1 **Priority algorithm**

Require: Data set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$; learning rate η; base classifier τ; total number of base classifiers T;

Ensure: Δy_i

1: **for** $m = 1 \rightarrow T$ **do**
2: $h_m = \tau(D_m)$
3: **end for**
4: Rebuilding data set pairs $\mathcal{P}$
5: Randomly initialize ω_m corresponding to h_m
6: **repeat**
7: **for** $all(p_i) \in \mathcal{P}$ **do**
8: $E_j = \xi(H(x_j), y_j)$
9: Calculate gradient $g = \frac{\partial \Delta y_i}{\partial \omega_m}$
10: $\Delta\omega_m = -\eta \cdot g$
11: Update ω_m
12: **end for**
13: **until** Achieving termination conditions
14: **return** $\Delta y_i = f(x_i) - f(x_{i+1})$

4.2 | Data imbalance

The experiment dataset suffers the problem of data label imbalance because compared to students without maladaptation, the number of students with maladaptation is less. SMOTE algorithm (Chawla et al., 2002) is used in our study to eliminate the bias caused by imbalanced data. Typical oversampling methods take a simple strategy that copies the target category sample, resulting in the law captured by learning models on modified data being too specific. By contrast, the basic idea of SMOTE is to analyze the categories with fewer data and to generate data by the following equation:

$$x_{new} = x + rand(0,1) \times (\tilde{x} - x), \tag{9}$$

where x and $\tilde{x}$ represent two different data in the same category. So, the learning model can capture more general patterns on modified data by the SMOTE algorithm.

5 | DATASET

The dataset used in this research includes 2.634 students from a Chinese university. These students come from 23 provinces and belong to 21 ethnic groups. They are

18 years old (mean = 18.825, $SD = 0.802$), and the ratio of males to females is 0.2:0.8. Our dataset contains about 456.000 records of data, including demographic data, psychological test data, and academic performance data. For ethics issues, all student data is securely stored in the data management department of the university they attend. As a research partner, we obtain the access right to anonymous data rather than raw data. All data we have access to is desensitized and does not contain any personally identifiable information relating to participants. The details are introduced as follows:

- Demographic data: For all universities, students are required to submit personal information at the time of admission, like hometown, gender, and ethnic group. Our dataset includes about 7,000 records of demographic data.
- Psychological test data: This research contains two psychological test scales, Eysenck personality questionnaire (EPQ) and SymptomChecklist90 (SCL 90), which are often used to assess students' psychological state (Wei et al., 2018; Yu et al., 2019). First, EPQ conceptualizes personality as four independent dimensions: E (extraversion/introversion), N (neuroticism/stability), P (psychoticism/socialization), and L (lie/social desirability). Because students who have scores of more than 60 in L scales are considered to have deliberately hidden the truth, we remove these students' data when calculating E, N, and P's coefficients. Second, the primary symptom dimensions of SCL-90 consist of total scores of psychological health (s1), somatization (s2), obsessive–compulsive (s3), interpersonal sensitivity (s4), depression (s5), anxiety (s6), hostility (s7), phobic anxiety (s8), paranoid ideation (s9), psychoticism (10), and a category of "additional scales" (s11) which helps clinicians assess other aspects of the clients' symptoms.
- Academic performance data: The dataset includes students' grades of college entrance examination scores and first-semester university exams, and it totally includes 1,048,575 records.

Note that, in the Chinese college entrance examination, each province designs its own test content and sets its own admission score. In other words, if the exam in Province A is easy, the scores of candidates in Province A will be higher, and the admission score will be higher. To be fair, we normalized the scores of candidates from different provinces based on the admission scores of their provinces. We standardize college entrance examination scores S_{real} according to admission scores $S_{\text{admission}}$. For student i, the standardization equation is as follows:

$$S_i = \frac{S_{\text{real}} - S_{\text{admission}}}{S_{\text{admission}}}. \tag{10}$$

6 | EXPERIMENTAL RESULTS

Our experimental results are displayed in this section. First, we analyze the correlation between three kinds of features (psychology, intelligence, demographic) and college freshman's maladaptation phenomenon. Second, the MASTER framework and the other four popular algorithms investigate the predictability of the maladaptation phenomenon. Since the prediction results are presented as a ranking, we use the Pearson correlation coefficient, Kendall correlation coefficient, and Spearman correlation coefficient to comprehensively evaluate the performance. Note that in order to avoid the bias caused by unique samples, each value in the article is the average result after 10 calculations.

6.1 | Maladaptation recognition

In this study, we utilize the null model method proposed by Cao et al. (2018) to define the threshold value of academic performance fluctuation for the recognition of the maladaptation phenomenon. We first construct a null model by randomly shuffling the record of R_{after}. For example, if one student's record of academic performance is $\{R_{\text{before}}, R_{\text{after}}\}$, we shuffle the record of R_{after}. Then we compute the mean and SD of academic performance backward by repeating the construction process 20 times. Second, we determine the threshold by keeping the performance backward of the real case. Above the mean performance backward plus two times of SD of the random case. Finally, the threshold is 0.3.

6.2 | Feature analysis

6.2.1 | Psychological feature analysis

As mentioned before, two widely accepted psychological questionnaires, EPQ and SCL-90, are involved in our study. We analyze the results in detail below.

First, the distribution of data collected through the EPQ questionnaire is shown in Figure 4a. This article does not use classical qualitative analysis of EPQ that plots the score on a two-dimensional coordinate axis to read testing participants' personalities. We use a quantitative method of analysis to analyze each score's correlation to the objectives of this study. The correlation coefficients between EPQ performance and academic performance

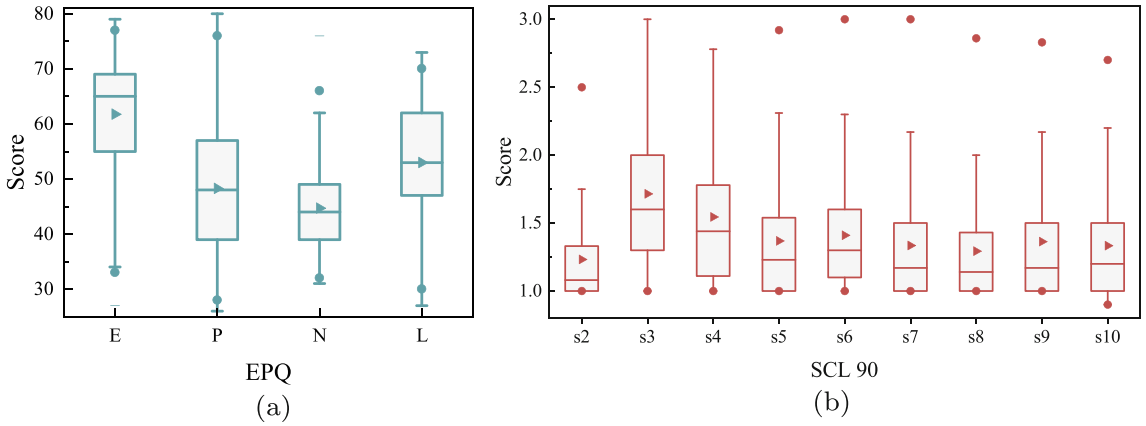

FIGURE 4 (a) Score distribution of EPQ. (Each scale, including *E*, *P*, *N* and *L*, is introduced in Section 5) (b) Score distribution of SCL-90. This box plot shows the distribution of the psychological tests scores of all the students participating in the experiment

TABLE 1 The correlation coefficient between EPQ and academic performance

	E	*N*	*P*	*L*
R_{before}	0.114**	−0.025	−0.022	0.048*
R_{after}	0.056	0.007	−0.086**	0.142**
α	−0.045*	0.027	−0.046*	0.071**

*$p < .05$; **$p < .01$.

are shown in Table 1. Note that the analysis results of *N* scale do not have a statistically significant correlation with R_{before}, R_{after}, and α, so only the results on the other three scales are analyzed.

E scale is characterized by being outgoing, talkative, high on positive affect (feeling good), and in need of external stimulation (Barrett et al., 1998; Eysenck & Eysenck, 1984). The higher the score, the more outgoing the personality. R_{before} and α all have a statistically significant correlation with the *E* scale. For R_{before}, diverse interests and flexible thinking result that this kind of student can accept basic knowledge from elementary and secondary education easily (Poropat, 2011). Moreover, current research demonstrated that compared to introverts, extroverts are more connected to their environment and more likely to amplify external stimulus (Argyle & Lu, 1990; Jung, 1921; Pavot et al., 1990). Based on this, we conjecture that introverts' insensitivity to the environment leads to the negative correlation between α and the *E* scale.

P scale is associated with the following traits: aggressive, assertive, egocentric, unsympathetic, manipulative, achievement-oriented, dogmatic, masculine, and tough-minded (Barrett et al., 1998; Eysenck & Eysenck, 1984). People with a high *P* scale are inclined toward being cold, impersonal, lacking in sympathy, antisocial, and lacking in insight (Heath & Martin, 1990). According to previous research (Francis, 1996; Francis & Montgomery, 1993), scoring on the *P* scale is inversely associated with academic performance, which is also consistent with our experiment result shown in Table 1. In addition, scholars have proved that the *P* scale in EPQ is related to the conscientiousness scale of the five-factor model, and the latter has substantial correlations with academic performance (Heaven et al., 2007). Another find here is that the *P* scale has a negative correlation with academic performance fluctuation α, and we conjecture that the reason for this result is that the negative state exhibited by the students with a high-scoring *P* scale makes it difficult for them to integrate into the new environment, which eventually leads to the decline of performance.

Moreover, *L* scale was originally introduced into EPQ to detect the "faking good" of scores on other scales (Jackson & Francis, 1998). However, scholars are aware that the *L* scale can reflect extra personality traits (Furnham, 1986; Jackson & Francis, 1998). According to the analysis results shown in Table 1, the *L* scale is the only one that has a positive correlation with two indicators of academic performance (i.e., R_{before} and R_{after}), which is consistent with the analysis results in existing studies (Chamorro-Premuzic & Furnham, 2003; Poropat, 2011). The positive correlation between academic performance fluctuation and scores of the *L* scale is consistent with the "positive attitude theory" (Poropat, 2011; White et al., 2008), which is that high scores on the *L* scale are associated with positive attitudes of classroom participation and related educational activities.

The other psychological scale used in this research is SCL 90 which is widely used by clinical psychologists, psychiatrists, and professionals in mental health, medical, and educational settings and research purposes

TABLE 2 The correlation coefficient between SCL 90 and academic performance

	s1	s2	s3	s4	s5	s6
R_{before}	−0.052**	−0.038	−0.029	−0.036	−0.047*	−0.052**
R_{after}	−0.020	−0.024	−0.027	0.003	−0.004	0.012
α	0.027	0.011	0.005	0.034	0.036	0.051**
	s7	**s8**	**s9**	**s10**	**s11**	
R_{before}	−0.033	−0.054**	−0.042*	−0.046*	−0.052**	
R_{after}	−0.030	0.020	−0.027	−0.024	−0.003	
α	0.004	0.058**	0.014	0.024	0.039*	

*$p < .05$; **$p < .01$.

(Preti et al., 2019; Tian et al., 2020). First, the distribution of data collected through the SCL 90 questionnaire is shown in Figure 4b. The SCL 90 includes 11 scales, including 10 psychological scales and the total score, and for the 10 psychological scales, each scale represents a type of negative psychological symptom. The higher the score, the more serious the negative psychological symptoms. According to the analysis results shown in Table 2, R_{before} has a statistically significant correlation with s1, s5, s6, s8, s9, s10, and s11. All statistical correlation is negative, which is consistent with common sense that negative psychological symptoms affect learning efficiency.

For academic performance fluctuation α, apart from s6, s8, and s11, other scales do not display statistically significant correlations. s6, s8, and s11 represent anxiety, phobic anxiety, and additional scales, and additional scales reflect the state of sleep and eating. These three scales are all related to stress (Allen et al., 2021; Hammen, 2005; Nojomi & Gharayee, 2007). In this case, we infer that positive stress stimulating motivation is the reason for the positive correlation between them (Park et al., 2012).

6.2.2 | Demographic feature analysis

We collect data from students' home provinces, administrative districts, ethnic groups, and genders for statistical analysis. As mentioned above, we classify students as follows: (a) We categorize students into two types according to their hometown and the province where their university is located: native students and non-native students. (b) Students are also divided into different categories according to administrative districts of their hometowns: town students and country students. (c) According to ethnic groups, we divide students into two categories: minority students and majority students. (d) All students are divided into male and female students. Then we calculate the percentage of students with the maladaptation phenomenon in each category, and the result is shown in Figure 5a.

First, we categorize students into native students and non-native students, based on the assumption that similar living habits will help them integrate into college life smoothly. It can be seen in Figure 5a that the percentage of students with the maladaptation phenomenon among native students is higher than the percentage of the non-native student ($p < .05$), which is inconsistent with our assumption. To verify this conclusion rigorously, we do further quantitative analysis. We calculate geographical linear distances between provincial capitals of students' hometowns and capitals of the province where the university is located, and calculate the correlation coefficient between geographical linear distances and academic performance fluctuation α. The geographical distribution of hometowns of the freshmen who participated in the experiment is shown in Figure 5b and, the relationship between geographical linear distances and academic performance fluctuation α is shown in Figure 6. The correlation coefficient between them is .254. This result demonstrates that the academic performance of students whose hometown is far away from school is prone to positive fluctuations in the first semester of college, which is consistent with the above results. We conjecture that this is related to the exam-oriented high school education system in China and the definition of maladaptation in this article (i.e., academic performance fluctuation). As Kirkpatrick and Zang (2011) mentioned in their research, high school students in China start school at 7:00 a.m. every morning. At school, each class is 45 min, with a 10-min break until 18:00 p.m. On average, there are 3 or 4 tests per subject, and there are very few extracurricular activities or hobbies due to the high volume of homework each day. Meanwhile, to let high school students concentrate on their studies, they are indoctrinated with an idea that college life is open and unfettered and are encouraged to develop their hobbies in college life (Hu et al., 2021; Li & Prevatt, 2008). In this case, we make the following conjectures: driven by a long-suppressed desire, these college freshmen are eager to explore the world outside the campus and experience a diversified life. For

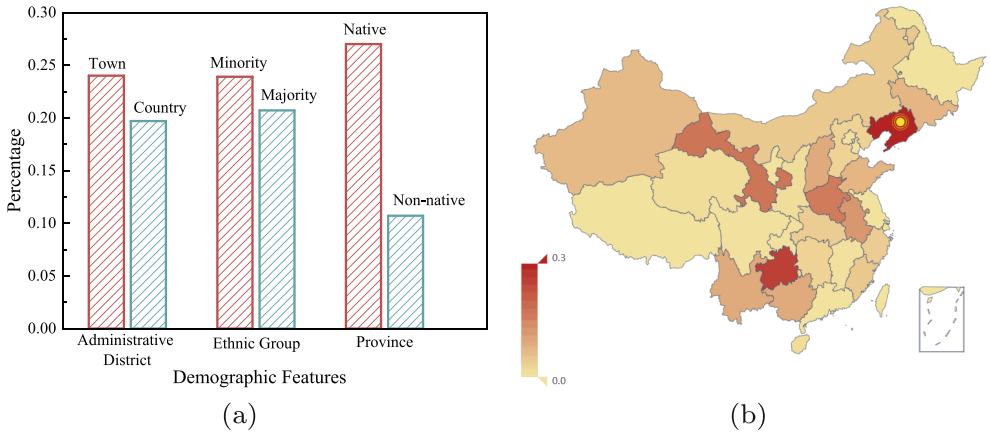

FIGURE 5 (a) The relation between demographic features and abnormal fluctuation of academic performance. (b) The geographical distribution of hometowns of the freshmen who participated in the experiment. The geographical distribution of hometown of the freshmen who participated in the experiment, the color represents its proportion. As the legend shows, the darker the color, the larger the proportion

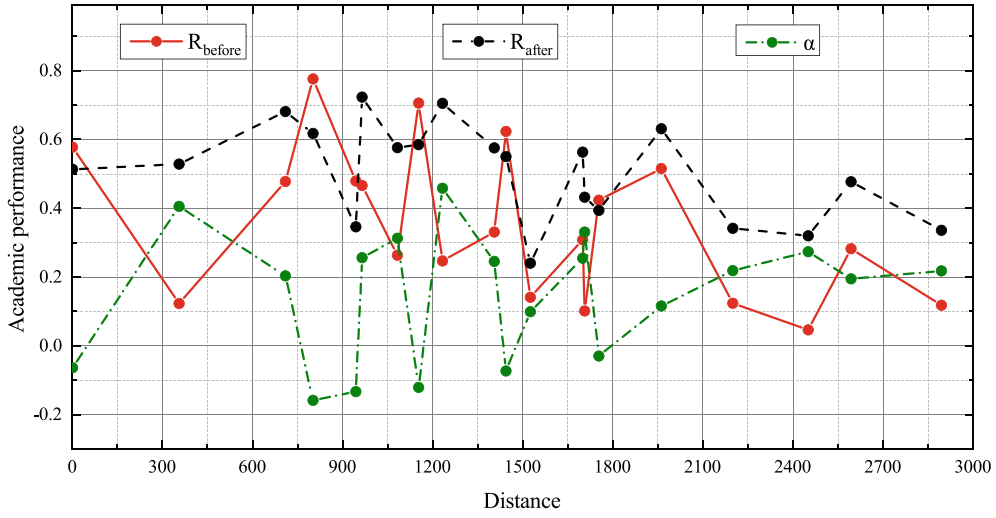

FIGURE 6 The relationship between geographical linear distances from hometown to the city where the school is located and students' academic performance. R_{before} is the performance of college entrance examinations used to represent the academic performance before starting university. R_{after} is the academic performance at the beginning of university life. α represents their difference

native students, familiarity with the environment and culture can make it all easier to start. In this case, they possibly have difficulty concentrating on studying. Note that the above is our conjecture about this anomalous result. We will explore this phenomenon more comprehensively in the following work.

Second, according to ethnic groups, we divide students into two categories: minority students and majority students. In Figure 5a, the percentage of students with maladaptation phenomenon among minority students is relatively higher than the rate of majority students, but this result is not statistically insignificant. This is related to the positive ethnic policy of the Chinese government (Rong, 2007; Wu, 2014), such as the bilingual education policy (L. Wang & Lehtomaki, 2021).

In addition, students are also divided into different categories according to the administrative districts of their hometowns: town students and country students. The result is shown in Figure 5a. Although the percentage of students with maladaptation among town students is higher than the country student, this result is not statistically insignificant due to $p > .05$. We conjecture that there are two reasons behind this phenomenon: First, over the past two decades, rapid economic development and the Chinese government's investment in infrastructure have narrowed the gap between urban and rural life (Y. Liu et al., 2016; Long et al., 2022). Second, the rapid development of the Internet makes the transmission of information no longer subject to geographical restrictions (DiMaggio et al., 2001). In short, there is no obvious difference between urban students and rural students in terms of living environment and information received.

Finally, all students are divided into male and female students according to their gender, and the analysis results are shown in Figure 7a. The percentage of male students with the maladaptation phenomenon is higher

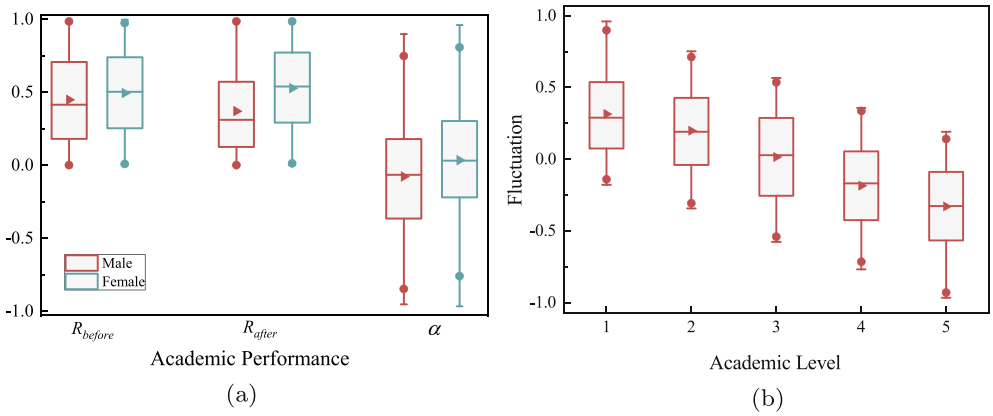

FIGURE 7 (a) The academic performance of male students and female students. (b) The relation between intelligence feature and academic fluctuation. According to R_{before}, students are divided into five levels from "5" to "1." The higher the level is, the better the R_{before} is

TABLE 3 The correlativity between each index of academic performance

	R_{before}	R_{after}	α
R_{before}	1		
R_{after}	0.166**	1	
α	−0.638**	0.637**	1

**$p < .01$.

than female students ($p < .05$). This is consistent with the conclusions drawn by other related scholars (Abdullah et al., 2009; Clinciu, 2013).

6.2.3 | Intelligence feature

As mentioned above, we explore how the intelligence feature influences the maladaptation phenomenon through statistically analyzing the relation between R_{before} and α. The correlation coefficients among academic performance indicators including R_{before}, R_{after} and α are displayed in Table 3. The correlation coefficient between R_{before} and R_{after} is 0.166, which represents that there is a positive correlation between them, but the degree is relatively weak. In other words, good performance in high school education cannot lead to the same result in higher education, proving the existence of the maladaptation phenomenon. The correlation coefficient between R_{before} and α is −0.638, which is related to the definition of the maladaptation in this research, so we will not analyze it here.

To further explore the detailed pattern behind the intelligence feature and academic fluctuation at the beginning of university life. We divide students into five academic levels according to R_{before} and the fluctuation of each level is shown in Figure 7b. It can be seen that the academic performance of students at different academic levels has varying degrees of fluctuation.

6.3 | Prediction

The significant correlations imply that the various features involved in our study could predict students' maladaptation. To explore the predictability, we experiment to predict whether students would suffer maladaptation based on the features that have been proved to display a statistically significant connection with the maladaptation phenomenon. We design a prediction framework named MASTER framework to overcome the challenges of such a special scenario. The whole experiment process is shown in Figure 2 and the detailed experiment set is shown as follows. Excluding some incomplete or error data records, our dataset includes 2,634 samples. We perform our prediction task as a score prediction experiment. We divide the dataset into two categories: students with maladaptation and students without maladaptation.

In this article, three different correlation coefficients are used to evaluate the prediction performance, and the details are shown as follows:

- Pearson correlation coefficient: Pearson's correlation coefficient is the test statistics that measures the linear relationship between two continuous variables, and the details are shown as follow:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{11}$$

- Kendall rank correlation coefficient: Kendall rank correlation coefficient is a nonparametric measure of relationships between ranked data columns.

$$Tau = \frac{C-D}{\sqrt{(N3-N1)(N3-N2)}}. \tag{12}$$

- Spearman correlation coefficient: The Spearman's rank coefficient of correlation is a nonparametric measure of rank correlation.

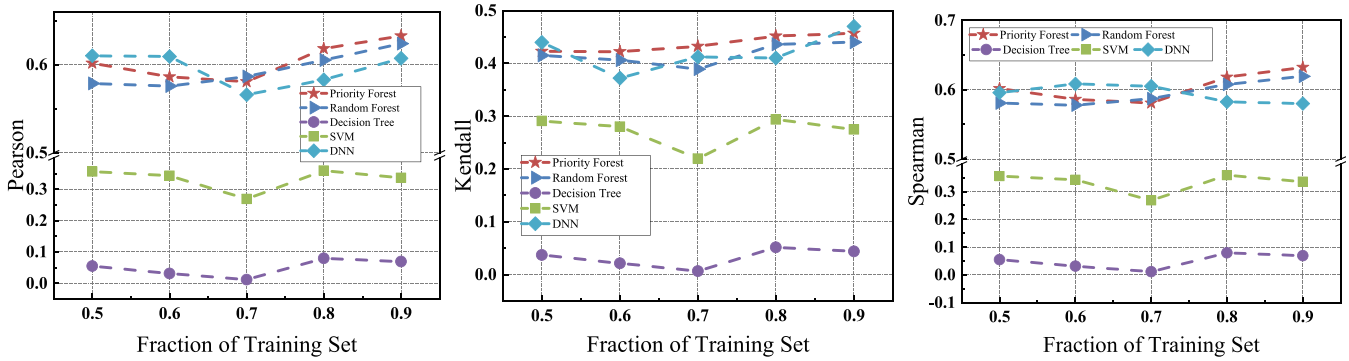

FIGURE 8 The results of the first-step experiment

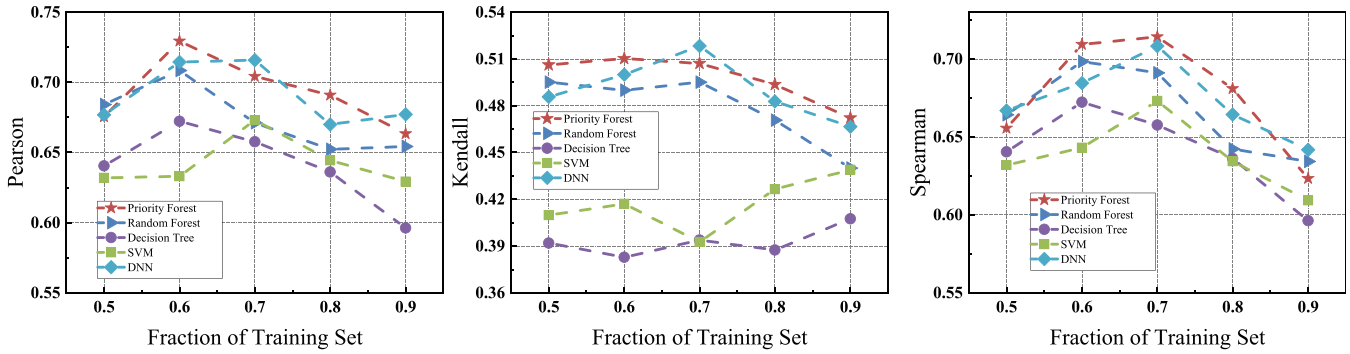

FIGURE 9 The results of the second-step experiment

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}. \tag{13}$$

To effectively train the model, we divided the data into three parts: training set, validation set, and test set. The training set is used to train each base classifier and the validation set is used to assign them weights. The final performance is evaluated through the test set. For each prediction task, we first remove 10%, 20%, 30%, 40%, and 50% of participants randomly from the dataset by stratified sampling to form a testing set, respectively, and we use the rest of the data as a training set.

To verify the effectiveness of our framework, we compare the MASTER framework with several popular algorithms shown as follows:

- Random forest: Random forest is a flexible and popular ensemble algorithm and is widely used in various fields (Bartley et al., 2019).
- SVM: SVM is a classic classification algorithm and is widely used in data mining (Victor et al., 2020).
- Decision tree: Decision tree is the most powerful and popular tool for classification (Z. Liu et al., 2020).
- DNN: DNN is a trendy model based on a multi-layer neural network and is widely used in various scenarios (LeCun et al., 2015). (Our research implements a common three-layer neural network model).

In the first step, we use raw data to fit all models and the results are shown in Figure 8. The imbalance label issue exists in raw data, leading to unexpected results that all evaluation indexes are relatively low. Even so, the priority forest proposed in this article performs better than the others.

In the second step, we carry out the second experiment that SMOTE is only used to solve the label imbalance on the training set. The process is shown as follows:

- First, raw data is divided into two categories: the training set $\boldsymbol{a}$ and testing set $\boldsymbol{b}$ by stratified sampling.
- Second, we use SMOTE on the training set $\boldsymbol{a}$ to generate samples of the minority class. Then in the new training set $\boldsymbol{a}'$, the number of students in the two classes is equal.
- Finally, we train our model and several popular algorithms on new training set $\boldsymbol{a}'$ and test their performances on the testing set $\boldsymbol{b}$.

The results are shown in Figure 9. From the above experimental results, we demonstrate the superiority of

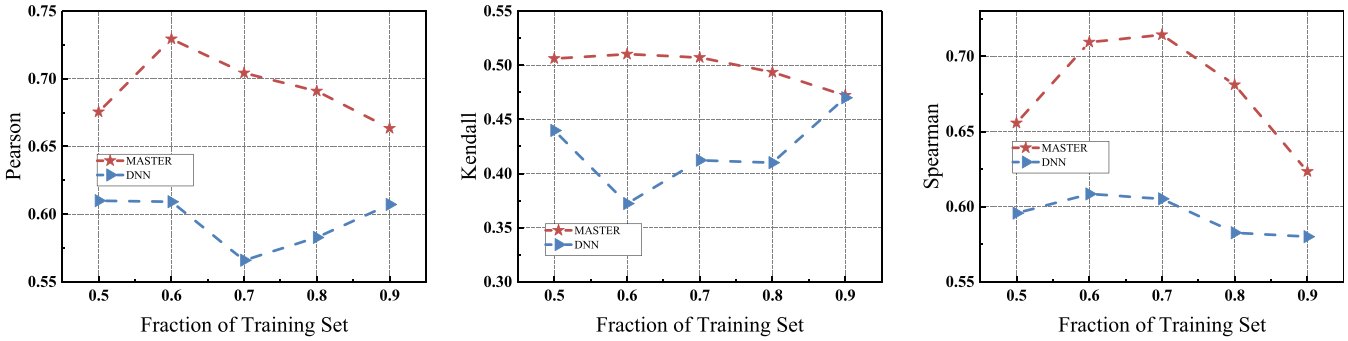

FIGURE 10 Comparison of the results of MASTER framework and DNN

the algorithm in multiple scenarios. Although the performance of DNN is very close to the algorithm proposed in this article, the inexplicability of DNN makes it difficult to gain the trust of all stakeholders and implement it (T. Guo et al., 2021; Xu, 2020).

Finally, to clarify how the MASTER framework performs, we compare the performance of the MASTER framework (SMOTE + priority forest) to DNN performance, which is the best in comparison experiments. The results are shown in Figure 10. In a word, all experiments above demonstrate the validity of our proposed framework.

7 | DISCUSSION AND CONCLUSION

This article analyzes the maladaptation phenomenon of college freshmen from three perspectives: intelligence features, psychological features, and demographic features. The maladaptation phenomenon is quantified through abnormal fluctuation of academic performance at the beginning of college life. In the meantime, our proposed MASTER framework is applied to investigate such a phenomenon's predictability. Experimental results show that the features mentioned above significantly correlate with the maladaptation phenomenon and prove its predictability. Additionally, the MASTER framework is verified that can improve prediction performance dramatically. To the best of our knowledge, we are the first to explore the maladaptation phenomenon using machine learning technologies on educational big data, which offers insights on a combination of psychological and machine learning techniques.

However, there are some limitations to this work. First, our research is limited by the unbalanced ratio of men to women in our dataset because the university involved in this experiment is a normal college. Second, although the SMOTE algorithm integrated into the MASTER framework is employed to ease the problem of data label imbalance effectively, the availability of balanceable labeled data is bound to be the most fundamental and effective solution. In view of the above two points, the most fundamental solution is to collect an open high-quality dataset. In the era of big data, a high-quality data set is a necessary condition to promote the development of the field. This will not only attract the attention of more researchers, but also get more reliable experimental results. However, this is not easy and requires the joint efforts of all stakeholders. Moreover, the recognition of maladaptation is just according to the decline of academic performance in this work even though it is the main indicator for students' performance, as many as possible factors should be considered for the comprehensive assessment of maladaptation, such as social situation.

There are multiple avenues for future work. Based on the MASTER framework, we plan to design an early warning system deployed in universities to provide references for policy makers. Meanwhile, we improve the performance of our framework according to their feedback. Second, although we use the fluctuations in academic performance to identify students with maladaptation and conduct experiments based on the Chinese dataset, we have proposed a broadly applicable idea, which is to detect whether students are adapting to university life by detecting changes in certain behaviors or characteristics before and after college entrance. In the future, we intend to define the maladaptation phenomenon from various aspects and explore the influence of maladaptation from various angles like the career plan and job hunting by the multi-dimension comparison of the adaptive and maladaptive.

ACKNOWLEDGMENT

Open access publishing facilitated by Federation University Australia, as part of the Wiley - Federation University Australia agreement via the Council of Australian University Librarians.

ORCID

Feng Xia https://orcid.org/0000-0002-8324-1859

REFERENCES

Abdullah, M. C., Elias, H., Mahyuddin, R., & Uli, J. (2009). Adjustment amongst first year students in Malaysian university. *European Journal of Social Sciences*, *3*, 496–505.

Allen, H. K., Barrall, A. L., Vincent, K. B., & Arria, A. M. (2021). Stress and burnout among graduate students: Moderation by sleep duration and quality. *International Journal of Behavioral Medicine*, *28*(1), 21–28.

Argyle, M., & Lu, L. (1990). The happiness of extraverts. *Personality and Individual Differences*, *11*(10), 1011–1017.

Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., Demyttenaere, K., Ebert, D. D., Green, J. G., Hasking, P., Murray, E., Nock, M. K., Pinder-Amaker, S., Sampson, N. A., Stein, D. J., Vilagut, G., Zaslavsky, A. M., Kessler, R. C., & WHO WMH-ICS Collaborators. (2018). Who world mental health surveys international college student project: Prevalence and distribution of mental disorders. *Journal of Abnormal Psychology*, *127*(7), 623–638.

Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A., & Xia, F. (2021). Educational big data: Predictions, applications and challenges. *Big Data Research*, *26*, 100270.

Barrett, P. T., Petrides, K. V., Eysenck, S. B., & Eysenck, H. J. (1998). The Eysenck personality questionnaire: An examination of the factorial similarity of *P*, *E*, *N*, and *L* across 34 countries. *Personality and Individual Differences*, *25*(5), 805–819.

Bartley, C., Liu, W., & Reynolds, M. (2019). Enhanced random forest algorithms for partially monotone ordinal classification. In *The thirty-third AAAI conference on artificial intelligence* (Vol. 33, pp. 3224–3231). AAAI Press.

Bernard, S., Adam, S., & Heutte, L. (2012). Dynamic random forests. *Pattern Recognition Letters*, *33*(12), 1580–1586.

Beyers, W., & Goossens, L. (2002). Concurrent and predictive validity of the student adaptation to college questionnaire in a sample of european freshman students. *Educational and Psychological Measurement*, *62*(3), 527–538.

Bonissone, P., Cadenas, J. M., Garrido, M. C., & Díaz-Valladares, R. A. (2010). A fuzzy random forest. *International Journal of Approximate Reasoning*, *51*(7), 729–747.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bruffaerts, R., Mortier, P., Auerbach, R. P., Alonso, J., Hermosillo De la Torre, A. E., Cuijpers, P., Demyttenaere, K., Ebert, D. D., Green, J. G., Hasking, P., Stein, D. J., Ennis, E., Nock, M. K., Pinder-Amaker, S., Sampson, N. A., Vilagut, G., Zaslavsky, A. M., Kessler, R. C., & WHO WMH-ICS Collaborators. (2019). Lifetime and 12-month treatment for mental disorders and suicidal thoughts and behaviors among first year college students. *International Journal of Methods in Psychiatric Research*, *28*(2), e1764.

Bruffaerts, R., Mortier, P., Kiekens, G., Auerbach, R. P., Cuijpers, P., Demyttenaere, K., Green, J. G., Nock, M. K., & Kessler, R. C. (2018). Mental health problems in college freshmen: Prevalence and academic functioning. *Journal of Affective Disorders*, *225*, 97–103.

Burrows, R. (2018). Understanding self-assessment in undergraduate dental education. *British Dental Journal*, *224*(11), 897–900.

Butler, S. M., Black, D. R., Blue, C. L., & Gretebeck, R. J. (2004). Change in diet, physical activity, and body weight in female college freshman. *American Journal of Health Behavior*, *28*(1), 24–32.

Cao, Y., Gao, J., Lian, D., Rong, Z., Shi, J., Wang, Q., Wu, Y., Yao, H., & Zhou, T. (2018). Orderliness predicts academic performance: Behavioural analysis on campus lifestyle. *Journal of the Royal Society Interface*, *15*(146), 20180210.

Chamorro-Premuzic, T., & Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality*, *37*(4), 319–338.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Y., Liu, Q., Huang, Z., Wu, L., Chen, E., Wu, R., Su, Y., & Hu, G. (2017). Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 989–998). ACM.

Clark, M., & Cundiff, N. L. (2011). Assessing the effectiveness of a college freshman seminar using propensity score adjustments. *Research in Higher Education*, *52*(6), 616–639.

Clinciu, A. I. (2013). Adaptation and stress for the first year university students. *Procedia-Social and Behavioral Sciences*, *78*, 718–722.

Credé, M., & Niehorster, S. (2012). Adjustment to college as measured by the student adaptation to college questionnaire: A quantitative review of its structure and relationships with correlates and consequences. *Educational Psychology Review*, *24*(1), 133–165.

Cui, Z., Chen, W., He, Y., & Chen, Y. (2015). Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 179–188). ACM.

De Vos, P., Hanck, C., Neisingh, M., Prak, D., Groen, H., & Faas, M. M. (2015). Weight gain in freshman college students and perceived health. *Preventive Medicine Reports*, *2*, 229–234.

Deshpande, S., Basil, M. D., & Basil, D. Z. (2009). Factors influencing healthy eating habits among college students: An application of the health belief model. *Health Marketing Quarterly*, *26*(2), 145–164.

DiMaggio, P., Hargittai, E., Neuman, W. R., & Robinson, J. P. (2001). Social implications of the internet. *Annual Review of Sociology*, *27*(1), 307–336.

Eisenberg, L. (1970). Student unrest: Sources and consequences. *Science*, *167*(3926), 1688–1692.

Eysenck, H., & Eysenck, S. (1984). *Eysenck personality questionnaire—Revised*. John Wiley & Sons.

Francis, L. J. (1996). The relationship between Eysenck's personality factors and attitude towards substance use among 13–15-year-olds. *Personality and Individual Differences*, *21*(5), 633–640.

Francis, L. J., & Montgomery, A. (1993). Personality and school-related attitudes among 11- to 16-year-old years. *Personality and Individual Differences*, *14*(5), 647–654.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*(3), 385–400.

Gewin, V. (2018). Depression tracker. *Nature*, *560*(7719), 519–520.

Gieseke, F., & Igel, C. (2018). *Training big random forests with little resources* (pp. 1445–1454). ACM.

Graetz, N., Friedman, J., Osgood-Zimmerman, A., Burstein, R., Biehl, M. H., Shields, C., Mosser, J. F., Casey, D. C.,

Deshpande, A., Earl, L., Reiner, R. C., Ray, S. E., Fullman, N., Levine, A. J., Stubbs, R. W., Mayala, B. K., Longbottom, J., Browne, A. J., Bhatt, S., ... Hay, S. I. (2018). Mapping local variation in educational attainment across africa. *Nature*, *555*(7694), 48–53.

Gravini Donado, M. L., Mercado-Peñaloza, M., & Dominguez-Lara, S. (2021). College adaptation among colombian freshmen students: Internal structure of the student adaptation to college questionnaire (SACQ). *Journal of New Approaches in Educational Research*, *10*(2), 251–263.

Gu, W., & Usinger, J. (2021). Independent learning, friendships, and fate in intercultural adaptation among international Chinese graduate students in the United States. *Journal of College Student Development*, *62*(1), 107–112.

Guo, H., Li, Y., Jennifer, S., Gu, M., Huang, Y., & Gong, B. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239.

Guo, T., Bai, X., Tian, X., Firmin, S., & Xia, F. (2021). Educational anomaly analytics: Features, methods, and challenges. *Frontiers in Big Data*, *4*, 811840.

Guo, T., Xia, F., Zhen, S., Bai, X., Zhang, D., Liu, Z., & Tang, J. (2020). Graduate employment prediction with bias. In *Proceedings of the 32th AAAI conference on artificial intelligence* (pp. 670–677). AAAI Press.

Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. *International Journal of Information Technology*, *12*(4), 1243–1257.

Hammen, C. (2005). Stress and depression. *Annual Review of Clinical Psychology*, *1*, 293–319.

Heath, A., & Martin, N. (1990). Psychoticism as a dimension of personality: A multivariate genetic test of Eysenck and Eysenck's psychoticism construct. *Journal of Personality and Social Psychology*, *58*(1), 111–121.

Heaven, P. C., Ciarrochi, J., & Vialle, W. (2007). Conscientiousness and Eysenckian psychoticism as predictors of school grades: A one-year longitudinal study. *Personality and Individual Differences*, *42*(3), 535–546.

Horgan, A., Sweeney, J., Behan, L., & McCarthy, G. (2016). Depressive symptoms, college adjustment and peer support among undergraduate nursing and midwifery students. *Journal of Advanced Nursing*, *72*(12), 3081–3092.

Hu, A., Wu, X., & Chen, T. (2021). Changing subjective wellbeing across the college life: Survey evidence from China. *Chinese Sociological Review*, *53*(4), 409–429.

Ikui, Y. (2019). A longitudinal study of school satisfaction: Mental health and social skills upon college admission. *Psychological Applications and Trends*, *1*, 255.

Jackson, C. J., & Francis, L. J. (1998). Interpreting the correlation between neuroticism and lie scale scores. *Personality and Individual Differences*, *26*(1), 59–63.

Johnson, T. P., & Smith, T. W. (2017). Big data and survey research: Supplement or substitute? In *Seeing cities through big data* (pp. 113–125). Springer.

Jung, C. G. (1921). *Psychologische typen*. Rascher.

Kim, H. B., Choi, S., Kim, B., & Pop-Eleches, C. (2018). The role of education interventions in improving economic rationality. *Science*, *362*(6410), 83–86.

Kirkpatrick, R., & Zang, Y. (2011). The negative influences of exam-oriented education on chinese high school students: Backwash from classroom to child. *Language Testing in Asia*, *1*(3), 1–10.

Krajniak, M. I., Pievsky, M., Eisen, A. R., & McGrath, R. E. (2018). The relationship between personality disorder traits, emotional intelligence, and college adjustment. *Journal of Clinical Psychology*, *74*(7), 1160–1173.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, C., Sung, Y.-T., Zhou, Y., & Lee, S. (2018). The relationships between the seriousness of leisure activities, social support and school adaptation among asian international students in the us. *Leisure Studies*, *37*(2), 197–210.

Li, H., & Prevatt, F. (2008). Fears and related anxieties in chinese high school students. *School Psychology International*, *29*(1), 89–104.

Liang, J., Zhang, X., Wang, J., Feng, L., Xu, C., Cheng, K., Cao, G., Yan, D., & Liu, B. (2020). Mental health status of college freshmen and influencing factors. *Psychology*, *11*(5), 737–747.

Liu, Y., Long, H., Chen, Y., Wang, J., Li, Y., Li, Y., Yang, Y., & Zhou, Y. (2016). Progress of research on urban–rural transformation and rural development in China in the past decade and future prospects. *Journal of Geographical Sciences*, *26*(8), 1117–1132.

Liu, Z., Wen, T., Sun, W., & Zhang, Q. (2020). Semi-supervised self-training feature weighted clustering decision tree and random forest. *IEEE Access*, *8*, 128337–128348.

Long, H., Ma, L., Zhang, Y., & Qu, L. (2022). Multifunctional rural development in China: Pattern, process and mechanism. *Habitat International*, *121*, 102530.

Nojomi, M., & Gharayee, B. (2007). Medical students and mental health by SCL-90-R. *Medical Journal of the Islamic Republic of Iran (MJIRI)*, *21*(2), 71–78.

Oh, C. Y., & Butler, B. (2019). Small worlds in a distant land: International newcomer students' local information behaviors in unfamiliar environments. *Journal of the Association for Information Science and Technology*, *70*(10), 1060–1073.

Panich, O., Tkachenko, N., Khudaeva, M., Ovsyanikova, Y., Sckilev, S., & Doronina, N. (2021). Cultural assimilator as a technology for preventing maladaptation of foreign students. In *SHS web of conferences* (Vol. 97, p. 01034). EDP Sciences.

Park, J., Chung, S., An, H., Park, S., Lee, C., Kim, S. Y., Lee, J.-D., & Kim, K.-S. (2012). A structural model of stress, motivation, and academic performance in medical students. *Psychiatry Investigation*, *9*(2), 143.

Patton, G. C., Olsson, C. A., Skirbekk, V., Saffery, R., Wlodek, M. E., Azzopardi, P. S., Stonawski, M., Rasmussen, B., Spry, E., Francis, K., Bhutta, Z. A., Kassebaum, N. J., Mokdad, A. H., Murray, C. J. L., Prentice, A. M., Reavley, N., Sheehan, P., Sweeny, K., Viner, R. M., & Sawyer, S. M. (2018). Adolescence and the next generation. *Nature*, *554*(7693), 458–466.

Pavot, W., Diener, E., & Fujita, F. (1990). Extraversion and happiness. *Personality and Individual Differences*, *11*(12), 1299–1306.

Poropat, A. E. (2011). The Eysenckian personality factors and their correlations with academic performance. *British Journal of Educational Psychology*, *81*(1), 41–58.

Preti, A., Carta, M. G., & Petretto, D. R. (2019). Factor structure models of the SCL-90-R: Replicability across community samples of adolescents. *Psychiatry Research*, *272*, 491–498.

Puff, J., Kolomeyer, E., McSwiggan, M., Pearte, C., Lauer, B.-A., & Renk, K. (2016). Depression as a mediator in the relationship between perceived familial criticism and college adaptation. *Journal of American College Health*, *64*(8), 604–612.

Resende, P. A. A., & Drummond, A. C. (2018). A survey of random forest based methods for intrusion detection systems. *ACM Computing Surveys (CSUR)*, *51*(3), 1–36.

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1619–1630.

Rong, M. (2007). Bilingual education for China's ethnic minorities. *Chinese Education & Society*, *40*(2), 9–25.

Sahão, F. T., & Kienen, N. (2021). University student adaptation and mental health: A systematic review of literature. *Psicologia Escolar e Educacional*, *25*, 1–9.

Sathe, S., & Aggarwal, C. C. (2017). Similarity forests. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 395–403). ACM.

Shaik, A. B., & Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model. In *International conference on innovative computing and communications* (pp. 253–260). Springer.

Soledad, R. G. M., Carolina, T. V., Adelina, G. C. M., Fernández, P., & Fernanda, M. (2012). The student adaptation to college questionnaire (SACQ) for use with spanish students. *Psychological Reports*, *111*(2), 624–640.

Speiser, J. L., Wolf, B. J., Chung, D., Karvellas, C. J., Koch, D. G., & Durkalski, V. L. (2019). Bimm forest: A random forest method for modeling clustered and longitudinal binary outcomes. *Chemometrics and Intelligent Laboratory Systems*, *185*(12), 122–134.

Tian, F., Li, H., Tian, S., Yang, J., Shao, J., & Tian, C. (2020). Psychological symptoms of ordinary Chinese citizens based on SCL-90 during the level I emergency response to COVID-19. *Psychiatry Research*, *288*, 112992.

Victor, B., Alberto, J., & Justo, P. (2020). Optimal arrangements of hyperplanes for SVM-based multiclass classification. *Advances in Data Analysis & Classification*, *14*(1), 175–199.

Wang, L., & Lehtomaki, E. (2021). Bilingual education and beyond: How school settings shape the Chinese Yi minority's sociocultural attachments. *International Journal of Bilingual Education and Bilingualism*, *25*, 1–13.

Wang, Y., Li, T., Noltemeyer, A., Wang, A., Zhang, J., & Shaw, K. (2018). Cross-cultural adaptation of international college students in the United States. *Journal of International Students*, *8*(2), 821–842.

Wei, Y., Li, H., Wang, H., Zhang, S., & Sun, Y. (2018). Psychological status of volunteers in a phase I clinical trial assessed by symptom checklist 90 (SCL-90) and Eysenck personality questionnaire (EPQ). *Medical science monitor: international medical journal of experimental and clinical research*, *24*, 4968–4973.

White, K. M., Thomas, I., Johnston, K. L., & Hyde, M. K. (2008). Predicting attendance at peer-assisted study sessions for statistics: Role identity and the theory of planned behavior. *The Journal of Social Psychology*, *148*(4), 473–492.

Wintre, M. G., Bowers, C., Gordner, N., & Lange, L. (2006). Re-evaluating the university attrition statistic: A longitudinal follow-up study. *Journal of Adolescent Research*, *21*(2), 111–132.

Wu, X. (2014). From assimilation to autonomy: Realizing ethnic minority rights in China's national autonomous regions. *Chinese Journal of International Law*, *13*(1), 55–90.

Xu, L. (2020). The dilemma and countermeasures of ai in educational application. In *2020 4th international conference on computer science and artificial intelligence* (pp. 289–294). ACM.

Yao, B., Han, W., Zeng, L., & Guo, X. (2013). Freshman year mental health symptoms and level of adaptation as predictors of internet addiction: A retrospective nested case–control study of male chinese college students. *Psychiatry Research*, *210*(2), 541–547.

Yu, Y., Wan, C., Huebner, E. S., Zhao, X., Zeng, W., & Shang, L. (2019). Psychometric properties of the symptom check list 90 (SCL-90) for Chinese undergraduate students. *Journal of Mental Health*, *28*(2), 213–219.

Zhang, D., Guo, T., Pan, H., Hou, J., Feng, Z., Yang, L., Lin, H., & Xia, F. (2019). Judging a book by its cover: The effect of facial perception on centrality in social networks. In *The world wide web conference* (pp. 2290–2300). ACM.

Zhou, Y., & Qiu, G. (2018). Random forest for label ranking. *Expert Systems with Applications*, *112*(12), 99–109.

How to cite this article: Guo, T., Bai, X., Zhen, S., Abid, S., & Xia, F. (2023). Lost at starting line: Predicting maladaptation of university freshmen based on educational big data. *Journal of the Association for Information Science and Technology*, *74*(1), 17–32. https://doi.org/10.1002/asi.24718