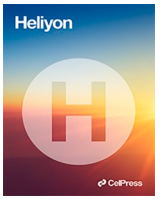
Heliyon 10 (2024) e29181

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A machine learning based method for constructing group profiles of university students

Ran Song [a,b], Fei Pang [c,*], Hongyun Jiang [a], Hancan Zhu [a,b]

[a] School of Mathematics, Physics and Information, Shaoxing University, Shaoxing, Zhejiang, 312000, China
[b] Institute of Artificial Intelligence, Shaoxing University, Shaoxing, Zhejiang, 312000, China
[c] Student Affairs Department, Shaoxing University, Shaoxing, Zhejiang, 312000, China

ARTICLE INFO

Keywords:
Questionnaire survey
K-means clustering
Group profiling
Neural network
Classification prediction

ABSTRACT

This study facilitates university student profiling by constructing a prediction model to forecast the classification of future students participating in a survey, thereby enhancing the utility and effectiveness of the questionnaire approach. In the context of the ongoing digital transformation of campuses, higher education institutions are increasingly prioritizing student educational development. This shift aligns with the maturation of big data technology, prompting scholars to focus on profiling university student education. While earlier research in this area, particularly foreign studies, focus on extracting data from specific learning contexts and often relied on single data sources, our study addresses these limitations. We employ a comprehensive approach, incorporating questionnaire surveys to capture a diverse array of student data. Considering various university student attributes, we create a holistic profile of the student population. Furthermore, we use clustering techniques to develop a categorical prediction model. In our clustering analysis, we employ the K-means algorithm to group student survey data. The results reveal four distinct student profiles: Diligent Learners, Earnest Individuals, Discerning Achievers, and Moral Advocates. These profiles are subsequently used to label student groups. For the classification task, we leverage these labels to establish a prediction model based on the Back Propagation neural network, with the goal of assigning students to their respective groups. Through meticulous model optimization, an impressive classification accuracy of 90.22% is achieved. Our research offers a novel perspective and serves as a valuable methodological reference for university student profiling.

1. Introduction

In recent years, university student profiling has experienced dynamic growth, both nationally and internationally [1]. However, a closer examination reveals several significant shortcomings that warrant attention. A key challenge lies in the limitations associated with the data sources. In pursuit of efficient data acquisition and processing, researchers have increasingly turned to questionnaire surveys as a valuable tool in educational informatics research [2]. Nonetheless, many studies still rely heavily on specific learning platforms or environments, resulting in a somewhat one-dimensional data collection that falls short of providing a comprehensive view of students' multifaceted attributes [3–5].

* Corresponding author.
E-mail address: pangf@usx.edu.cn (F. Pang).

https://doi.org/10.1016/j.heliyon.2024.e29181
Received 19 October 2023; Received in revised form 1 April 2024; Accepted 2 April 2024
Available online 7 April 2024

Furthermore, in terms of research methodologies, some scholars remain entrenched in traditional statistical techniques, neglecting the potential of emerging technologies such as machine learning. This approach limits the depth of insights obtained from complex student datasets [6]. Moreover, most studies overlook cross-cultural and cross-regional factors, casting doubt on the universality of student profiles across diverse cultural and geographical contexts. Additionally, many studies overly emphasize quantitative data, often neglecting the exploration of softer metrics such as students' psychology, emotions, and social interactions that could add a more profound humanistic dimension to their findings [7].

In summary, the field of student profiling requires refinement in several crucial areas, including data acquisition, research methodologies, and expansion of the dimensions of analysis. Our investigation delves into the construction and categorization of student educational profiles [8–10]. To provide beneficial insights and predictions in areas such as students' learning tasks, occupational effectiveness, and future contributions, this study offers a more valuable student analysis for institutions such as universities. We conduct an in-depth analysis of multidimensional data, including students' learning behaviors, hobbies, and social activities. Questionnaires based on the research questions are used to collect this diverse dataset. To achieve a differentiated group classification, the K-means clustering algorithm is employed to categorize the university student population into the following groups: diligent learners, morally upright individuals, discerning minds, and practical individuals. This process helps reveal the potential strengths and areas of interest for each student category, thereby better supporting educational institutions and students in selecting suitable academic disciplines and task domains. Furthermore, to rapidly assess the potential value of university students and facilitate strategic planning for training and learning programs by businesses and educational institutions, we construct a university student population classification prediction model based on a Back Propagation (BP) neural network using the K-means clustering results as a foundation. The model aims to accurately predict the categories of students who would complete future surveys. This predictive approach enables a swift understanding of the potential contributions and impacts that students may bring to their professional and academic careers, providing a tool for businesses, recruiting agencies, and higher education institutions to gain a profound insight into students' potential and value.

2. Related work

2.1. Group profiling techniques

The concept of user profiling revolves around the strategic utilization of advanced information technology to gather user data. By employing data mining techniques, this approach delves deeply into user characteristics, ultimately providing precise and detailed descriptions of various user attributes [11]. In the context of analyzing student data with heightened precision, adopting a user-profiling strategy is highly recommended for constructing a comprehensive profile of the university student population.

For instance, Constantinides et al. [12] harness interaction logs from news readers, and apply user modeling techniques to construct personalized user profiles. Meanwhile, Asif et al. [13] employ data-mining methodologies to investigate the academic performance of university students, resulting in the identification of two distinct student cohorts: those with lower and higher achievement. Yuan et al. [14] introduce the non-parametric Bayesian model EW4, which is specifically designed to model users' mobile behaviors and gain insights into user interests and intentions. These existing studies illustrate the diverse applications and benefits of user-profiling strategies in various domains, including education.

2.2. Classification prediction models

In the realm of data mining, although clustering methodologies hold a prominent place, classification techniques are equally vital to researchers. Clustering groups samples based on their similarities, whereas classification derives classification rules from known sample characteristics, creating decision formulas and discriminant criteria. In the context of classification prediction research, Shen et al. [15] comprehensively analyze comparing decision trees, neural networks, and logistic regression to examine the various types and mechanisms of credit card fraud. They evaluate the accuracy of fraud detection. Thomassey et al. [16] introduce a forecasting system that combines clustering and decision trees and assess its effectiveness using real-world data. Angiulli [17] introduce an innovative Nearest Neighbor Condensation algorithm, named FCNN, that specifically is designed for the classification of large datasets. Additionally, Şen et al. [18] utilize an extensive dataset from Turkey and implement a centralized placement test algorithm to identify factors correlated with future academic achievements. These existing studies illustrate the wide-ranging applications of classification techniques in diverse domains and highlight their significance in data mining research.

2.3. Student group profiling research

Within the domestic context, practical examples in which higher education institutions have applied user profiling techniques to create profiles for their university student cohorts are limited. Thus, a further investigation of this methodology is necessary. Researchers must use all university resources to systematically gather student data and extract the embedded latent characteristics. As university students continually strive for comprehensive personal development, higher education institutions must refine their service systems to guide students toward holistic growth. Group profiling techniques can categorize students with similar attributes into specific cohorts, thereby providing valuable insights for universities in areas such as assistance, mentorship, and support. These insights can positively influence decisions related to educational management [19]. Group-profiling techniques are expected to gain widespread acceptance in university management by simplifying the inherent complexities of the process. Looking ahead, researchers

must further enhance their models and strategies to optimize the construction of student cohort profiles. This ongoing refinement facilitates the more effective use of user profiling techniques in higher education institutions and ultimately benefit both students and the institutions themselves.

2.4. Design of the overall framework for this study

Fig. 1 illustrates the study's overall framework that comprises two tasks: clustering and classification. For the clustering task, we initially design questionnaires and collect multidimensional data from college students based on the classification objectives of the student population, considering multiple dimensions. Subsequently, we apply the K-means clustering algorithm to cluster the data and construct a comprehensive profile of the college student population based on various dimensional features. To swiftly predict the categories of college student groups, in the classification prediction task, we use clustering results as a foundation. We obtain category labels for the college student population and subsequently establish a BP neural network classification prediction model. This organic integration of clustering and classification tasks has significant implications for educational resource planning, student career guidance, improving university graduation rates, cultivating outstanding talents, and providing methodological frameworks for questionnaire surveys.

3. Methods of constructing university student group profiles

3.1. Design philosophy of group profiling

We meticulously outline the procedural steps for the development of university student cohort profiles, as illustrated in Fig. 2. The construction of the university student population profile is divided into three main stages: data collection, label generation, and visual presentation. First, in the datafication stage, we establish explicit research objectives centered on classifying the university student population. Then, during the questionnaire design phase, we carefully formulate several questions across multiple dimensions to capture valuable data from student responses and lay the foundation for a comprehensive feature architecture underpinning the cohort profiles. During the labeling stage, we harness a clustering model based on the K-means algorithm. By inputting the relevant data attributes, we successfully derive an optimal clustering outcome, identifying four distinct student groups. Building on this robust foundation, we leverage user-profiling technology to conduct an in-depth analysis. This analysis elucidates the unique attributes that characterized each of the four groups, culminating in the creation of distinct cohort profiles. Finally, in the visualization phase, we enhance the intuitive presentation of these profiles through a visual representation that underscores the pivotal feature distribution across the four profile categories. This systematic approach enables us to construct and visually represent comprehensive university student cohort profiles, offering valuable insights into the diverse attributes and characteristics of the student population.

3.2. Questionnaire design and data collection

3.2.1. Questionnaire design

To explore the university student population, we employ a questionnaire survey as the primary method of data collection. In terms of dimensional design, we draw inspiration from the role creation approach of Mulder and Yaar [20], which mainly includes the dimensions of "goals, behaviors, and attitudes" to explore user profiles. We use the individual cognition and experience of university students as the basis for segmentation, employing a three-dimensional perspective to delve into the study of user profiles to gain a more comprehensive understanding. This inspiration allows us to design multiple dimensions to explore potential information about the university student population more deeply. In the design of specific question content, we further reference the VALS2 model based on

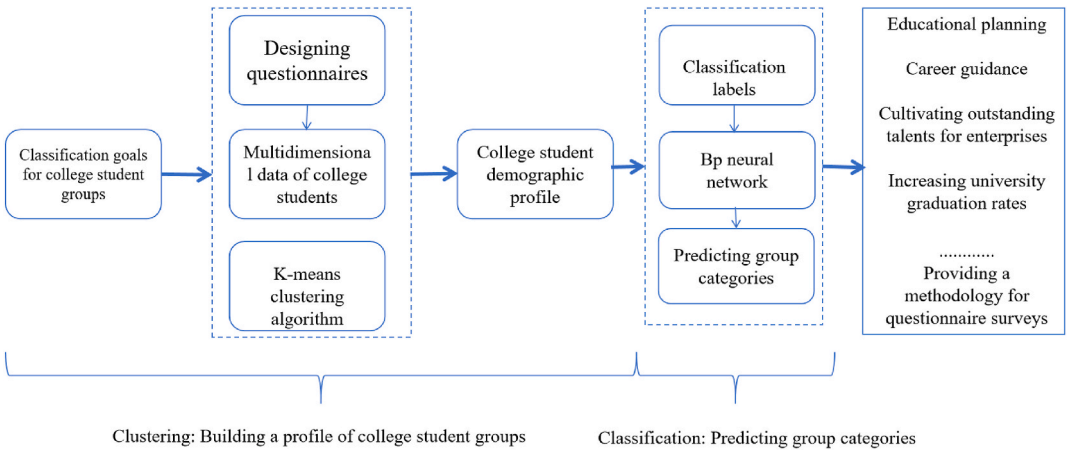


Fig. 1. Overall framework diagram.

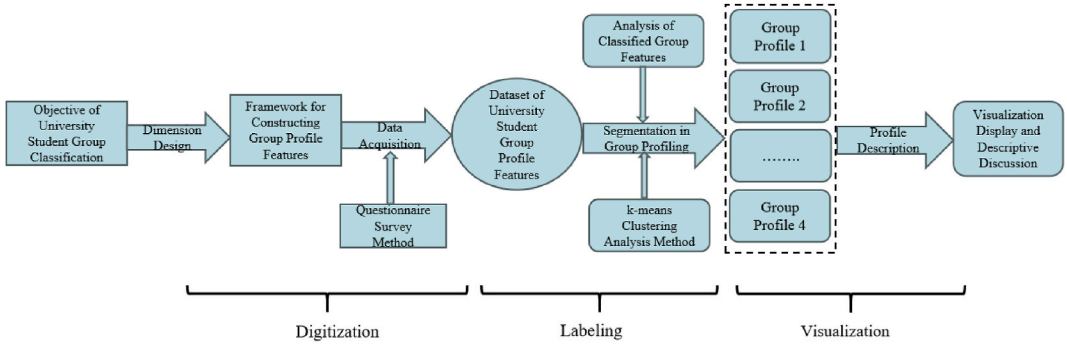


Fig. 2. Workflow diagram for group profiling construction.

Maslow's hierarchy of needs theory and motivation theory [21]. The VALS2 model is the mainstream method used for market segmentation. By borrowing from the VALS2 model, we consider factors such as user values, hobbies, behavior patterns, and attitude beliefs, and used them as references for constructing the user scale to ensure that our questionnaire content aligned closely with the characteristics and needs of the university student population. The design of such dimensions and questions allows for a more comprehensive understanding of the study, minimizing the duplication of previous research and maintaining a focused and unique perspective on the university student population.

3.2.2. Main components of the questionnaire

During the questionnaire design process, we devise a series of questions spanning multiple dimensions that enable a comprehensive exploration of students' multifaceted attributes and leverage data analysis tools to construct a portrait of the university student cohort. As illustrated in Fig. 3, the survey is divided into five sections. The first section encompasses the demographic details of the students,

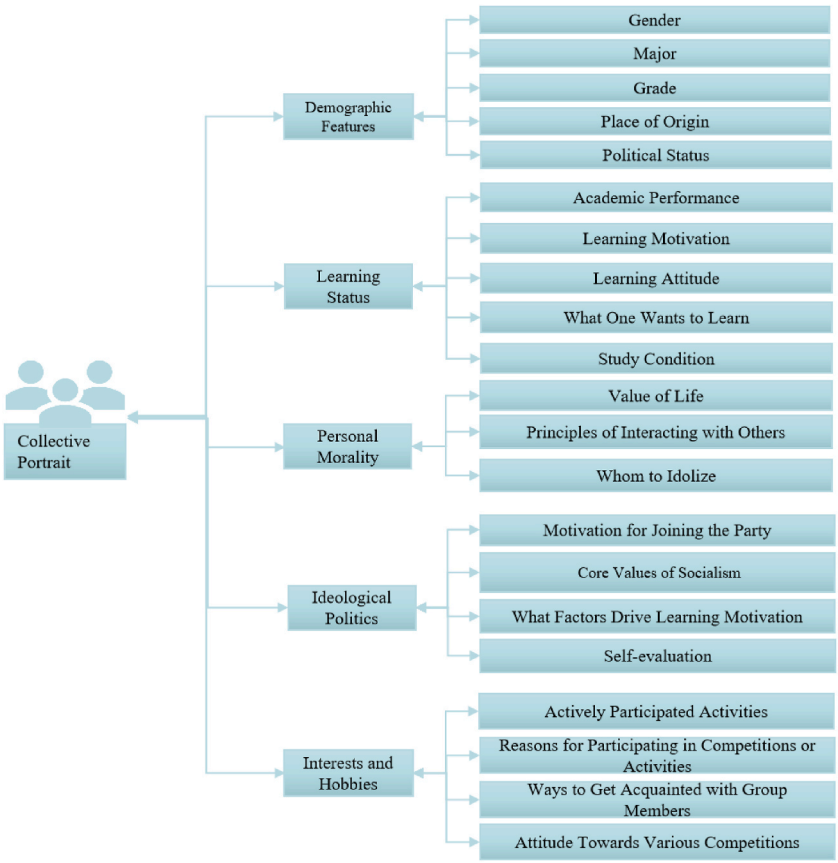


Fig. 3. Primary composition of the questionnaire.

including their gender, grade level, and study major. The second section focuses on academic circumstances and covers aspects such as academic performance and attitudes toward learning. The third section delves into students' values and qualities, incorporating their perspectives on life and attitudes toward team collaboration. The fourth section examines the students' political awareness, particularly their stance on the core values of socialism. Lastly, the fifth section focuses on students' interests and hobbies, encapsulating their motivations for participating in activities and modes of interaction with peers.

To frame the questions, we adopt a format akin to a Likert scale with a total of 22 items. The survey is crafted and disseminated through online platforms such as Wenjuanxing. Targeting university students as the respondent group, questionnaires are randomly distributed and accompanied by clear guidelines.

3.2.3. Demographic feature analysis

We successfully collect 2492 responses from university students. As shown in Table 1, the male-to-female ratio among the collected samples is approximately 1:1. The sample comprises students from various academic years, with first-year students constituting the predominant group. Furthermore, the collected samples span various academic disciplines, encompassing fields such as literature, management, education, natural sciences, engineering, and the arts. Notably, nearly half of the university students who participated in the survey are members of the Communist Youth League, whereas the proportion of those who are Communist Party members is relatively smaller.

3.2.4. Data preprocessing workflow

In the data preprocessing stage, we import the collected questionnaire data into the SPSS software for analysis and eliminate invalid data entries (such as incomplete questionnaires that were not filled out in their entirety). Given that the scales and magnitudes of the various variables within the data do not show significant discrepancies, we do not standardize the feature attributes of each sample's selected options.

3.3. Validity analysis of group profile survey results

The results reveal a balanced gender ratio among university students, with the samples encompassing a diverse range of academic years and major fields, including literature, natural sciences, engineering, and the arts. The extensive distribution characteristics, coupled with a large sample size (2492 samples) and the random distribution of questionnaires, contribute to the excellent representativeness of the study. To assess the questionnaire's validity and reliability, we utilize the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests. The KMO test shows a value of 0.838, which is higher than the commonly accepted threshold of 0.7. This indicates that our questionnaire is well-suited for factor analysis. Moreover, Bartlett's test result yields an approximate chi-square value of 11,688.101 with a significance level (p-value) of less than 0.001; thus, the null hypothesis is rejected. The result shows a significant correlation among the questionnaire items, suggesting their suitability for factor analysis and confirming the strong validity and reliability of our questionnaire's structure. Therefore, the robustness of the questionnaire's design is affirmed.

Table 1
Demographic variables statistics table.

Demographic Variables Frequency Analysis			
Name	Option	Frequency	Percentage
Gender	Male	1298	52%
	Female	1194	48%
Grade	Freshman	738	30%
	Sophomore	633	25%
	Junior	556	22%
	Senior	478	19%
	Others	87	4%
Major	Literature	268	11%
	Management	269	11%
	Education	362	15%
	Science	487	20%
	Engineering	351	14%
	Medicine	281	11%
	Arts	219	9%
	Others	255	10%
Student Leaders	Yes	1158	47%
	No	1334	54%
Only Child	Yes	1230	49%
	No	1262	51%
Place of Origin	City	1123	45%
	Rural	1369	55%
Political Status	Communist Party Member	107	4%
	Communist Party Activist	517	21%
	Communist Youth League Member	1137	46%
	General Public	731	29%

4. Clustering model

4.1. K-means clustering algorithm

4.1.1. Steps of the K-means clustering algorithm

We employ the K-means algorithm to cluster the gathered dataset [22,23]. The specific steps of the algorithm are as follows:

(1) k objects from the dataset are randomly selected to serve as the initial clustering centroids $c_i(i = 1,2,...k)$. Each centroid has d dimensions and is denoted as $c_{ij}(j = 1,2,...d)$.
(2) The Euclidean distance from each data point to the clustering centroids is computed. Based on the computed distances, each data point is assigned to the nearest centroid, resulting in the formation of new clusters.
(3) For each cluster, the centroid is recalculated by taking the mean value of all data points within that cluster, thereby determining the new centroid location.
(4) Steps 2 and 3 are repeated until the centroids no longer change or a predefined number of iterations is reached.

4.1.2. Determination of the optimal number of clusters

Numerous studies employ the elbow method to ascertain the optimal number of clusters. The pivotal metric of the elbow method is the Sum of Squared Errors (SSE), which represents the squared sum of the distances between each data point and its corresponding cluster center [24],

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$$

where C_i represents the i^{th} cluster, p is the sample point within cluster C_i, and m_i is the centroid of cluster C_i (the mean of all samples in C_i). The SSE signifies the clustering error of all samples and serves as an indicator of the quality of the clustering results.

We employ the K-means clustering method to categorize college students across four dimensions. By incrementally increasing the cluster count k, we refine the categorization of the collected questionnaire samples, thereby enhancing the degree of intra-cluster cohesion. However, when k is less than the actual number of clusters, augmenting k significantly bolsters the cohesion within the clusters, resulting in a sharp decline in the SSE [25,26]. However, once k reaches the true number of clusters, the benefit of increasing k gradually diminishes and the rate of the SSE reduction decelerates, ultimately stabilizing. Hence, we utilize the elbow method to determine the optimal number of clusters, represented by the "elbow" point, as depicted in Fig. 4. We observe a pronounced enhancement in clustering efficacy when four or five clusters are selected. Through a discriminant analysis of the clustering outcomes, we ultimately determined that k = 4 is the most suitable cluster count.

4.2. Group profile feature analysis based on clustering algorithm

Based on the clustering results, we construct detailed profiles of the four distinct student groups and designated names for each group. Owing to the questionnaire's item structure, which employs a Likert scale format, the item values increase incrementally from 1 to 5, representing a gradient from mild to strong intensities. For instance, the three questions in Table 2 aim to investigate students' academic statuses: Question 13 evaluates academic performance, where 1 signifies below average and 5 indicates excellent; Question 14 measures academic motivation, where 1 indicates low motivation and 5 signifies high motivation; Question 15 appraises study attitudes, where 1 symbolizes a lack of seriousness and 5 represents utmost seriousness. Beyond the basic demographic data, we also statistically analyze the scores derived from the students' selections by calculating both the mean and standard deviation of the scores,

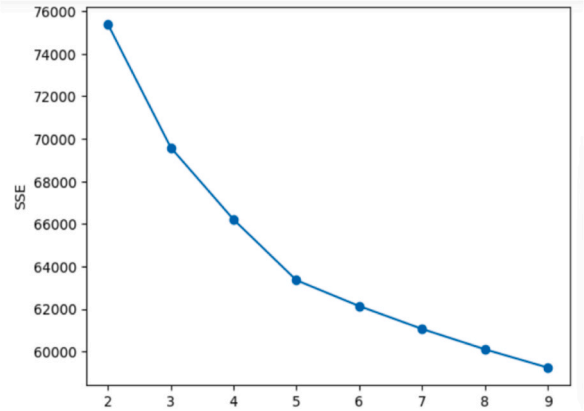


Fig. 4. Cluster analysis of group characteristics: SSE values for different k.

Table 2
Partial question topics and options.

Question	Option
13. Your current academic performance in your class	1 being poor, 5 being excellent 1 2 3 4 5
14. Your motivation for studying in university	1 being very weak, 5 being very strong 1 2 3 4 5
15. How do you evaluate your attitude towards studying?	1 being not serious at all, 5 being very serious. 1 2 3 4 5

thereby offering a profound description of the unique characteristics inherent to each group.

Through an in-depth statistical analysis of the scores associated with the options selected by the students, we compute the mean and standard deviation for each item. The specific statistical data are presented in Table 3. The following is an analysis of the characteristics of each clustered group accompanied by the designated nomenclature.

Cluster 1: Students in this group have a higher score in the academic performance dimension. Upon further examination of Question 22 within the interest and hobbies dimension, we discern that these students have a proclivity towards participating in academic contests and entrepreneurial innovation competitions. They are called "Diligent Learners."

Cluster 2: Although students in this category have average scores in the academic performance dimension, they excel in the personal character dimension. Moreover, a significant number of students from this cluster are keen to participate in volunteer services. They are called "Earnest Individuals."

Cluster 3: Students in this cluster have the highest scores across all four primary dimensions (academic performance, personal character, political thought, and interests and hobbies). This suggests that they possess commendable comprehensive qualities and have established a correct worldview, life philosophy, and set of values. They are called "Discerning Achievers."

Cluster 4: Students in this category consistently score higher in the personal character and political thought dimensions, indicating their emphasis on personal character cultivation, guided by socialist values, and heightened political awareness. They are called "Moral Advocates."

We distinguish the four student groups in terms of their significant characteristic disparities across multiple dimensions including academic performance, personal character, political ideology, and interests and hobbies. To present these inter-group differences more intuitively, we visualize the characteristics of each student group, and the specific visual representations are displayed in Table 4.

4.3. Practical use and significance of constructing group portraits of college students

Based on the classification of university student groups, we can provide educational planning guidance and career advice tailored to each cluster:

1. Diligent Learners:

Students should be encouraged to maintain their enthusiasm for learning and provided with more opportunities for subject competitions and innovation, and entrepreneurship competitions to broaden their perspectives and practical skills. Future career choices related to their subject competitions and innovation, and entrepreneurship competitions, such as fields in science and technology innovation, and engineering should be recommended. Additionally, active participation in internships and projects is advised to enhance practical skills.

2. Practical Individuals:

Table 3
Statistical mean and variance of option features.

Descriptive Statistics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Question 13	3.02 ± 1.428	3.02 ± 1.398	3.42 ± 1.161	3.06 ± 1.361
Question 14	2.99 ± 1.408	2.97 ± 1.426	3.55 ± 1.175	3.02 ± 1.375
Question 15	3.02 ± 1.432	2.94 ± 1.385	3.54 ± 1.169	2.98 ± 1.36
Question 17	2.49 ± 1.139	2.47 ± 1.112	3.11 ± 1.031	2.45 ± 1.113
Question 18	2.5 ± 1.113	2.55 ± 1.109	3.12 ± 0.905	2.54 ± 1.094
Question 9	3.01 ± 1.412	3.14 ± 1.416	4.06 ± 1.112	3.01 ± 1.415
Question 24	3.53 ± 1.634	2.85 ± 1.392	7.57 ± 1.305	7.55 ± 1.101
Question 33	5.89 ± 1.004	2.28 ± 1.026	6.11 ± 1.083	2.87 ± 1.556
Question 10	2.93 ± 1.382	3.13 ± 1.433	4.28 ± 1.127	3.16 ± 1.406
Question 21	4.05 ± 1.408	3.9 ± 1.38	4.81 ± 1.132	4 ± 1.402
Question 23	3.09 ± 1.432	3.04 ± 1.465	3.83 ± 1.371	3.02 ± 1.461
Question 25	2.95 ± 1.385	2.99 ± 1.399	2.27 ± 1.222	3.02 ± 1.404
Question 30	2.49 ± 1.12	2.48 ± 1.12	3.17 ± 0.955	2.47 ± 1.141

Table 4
Description and visualization of college students' user personas.

User Profile Types		Diligent Learning	Earnest Individuals	Discerning Achievers	Moral Advocates
Key Features		Passionate About Learning Excellent Academic Performance Strong Learning Motivation	Average Academic Performance Enthusiastic About Volunteering Activities Focus on Developing Hobbies and Interests	High Comprehensive Quality in Various Aspects	Emphasis on Personal Character Cultivation Guided by Socialist ValuesHigh Political Awareness
Gender Distribution	Male	49.4%	49.7%	57.7	51.4%
	Female	50.6%	50.3%	42.3%	48.6%
Grade Distribution	Freshman	24.8%	25.5%	39.6%	28.3%
	Sophomore	25.3%	24.1%	26.2%	25.9%
	Junior	24%	23.4%	19.6%	22.3%
	Senior	21%	22.7%	12.6%	20.6%
	Others	4.8%	4.3%	1.9%	3%
Major Distribution	Literature	12.6%	14.5%	5.6%	10.7%
	Management	13.1%	12.9%	6.2%	11.1%
	Education	13.5%	12.7%	16%	15.6%
	Science	12.1%	11.4%	36.7%	17.4%
	Engineering	11.1%	13.3%	20.8%	11.3%
	Medicine	12.9%	12.7%	5%	14.4%
	Arts	11.3%	10.8%	4.5%	8.7%
	Others	13.4%	11.7%	5.3%	10.7%
Place of Origin Distribution	City	46.1%	52.3%	35%	47.2%
	Rural	53.9%	47.7%	65%	52.8%
Political Affiliation Distribution	Political Status	3.7%	4.5%	3.7%	5.3%
	Communist Party Member	23.4%	23.2%	16%	20.6%
	Communist Youth League Member	43.2%	43.4%	49%	46.6%
	General Public	29.7%	28.9%	31.3%	27.5%

Educational institutions can emphasize the importance of personal character and volunteer services, encouraging participation in social activities beyond academics to cultivate practical problem-solving skills. Career paths in social services, social work, and nonprofit organizations as well as positions in management and leadership should be recommended to leverage personal character and leadership strengths.

3. Discerning Minds:

They should be provided with more in-depth subject knowledge and interdisciplinary learning opportunities to encourage in-depth research in multiple fields. Career choices in areas with higher demands for comprehensive skills, such as law, international relations, humanities, and social sciences should be recommended. Active participation in social activities is encouraged to become influential societal leaders.

4. Morally Upright Individuals:

Educational institutions should emphasize the cultivation of personal character and political ideology, encourage a deep understanding of socialist core values, and demonstrate a sense of social responsibility in their studies. Career choices aligned with socialist core values, such as political science and public administration should be recommended. Involvement in social affairs is encouraged to become agents of social change.

Therefore, although existing studies examine the educational profiles of university students, they primarily focus on mining learning context data, with some studies discussing aspects of students' social interactions. Additionally, existing studies often suffer from data singularity, clutter, and redundancy, resulting in insufficient comprehensiveness and accuracy when constructing educational profiles. In contrast, our innovative method for constructing university student group profiles provides a novel approach in terms of methodology and data acquisition, thus offering a multidimensional assessment of students and constructing a more comprehensive and detailed university student group profile.

5. Establishing a university student group classification prediction model

We provide a detailed description of how the student questionnaire selection is transformed into feature vectors. Using the K-means clustering method and attributes such as academic performance, interests, morality, and political thought, we categorize the university student population into four groups: Diligent Learner, Morality Cultivator, Discerning Thinker, and Practical Realist. Drawing on these clustering results, we construct a classification prediction model for university student groups based on the BP neural network [27–29].

The algorithm used in the model, namely the BP neural network, is a type of multi-layer feed forward network trained using the BP algorithm. From the clustering analysis, the student groups are segmented into the four categories. Correspondingly, we assign category labels to the 2492 samples. These four categories are numerically represented by the numbers 1, 2, 3, and 4.

For each sample, the result of a single-choice question is mapped to an element of the feature vector, as listed in Table 5. After determining the features for single-choice questions, we also incorporated multiple-choice questions into the feature vector in the following manner. Question 8 delves into the motivations for students to join the Communist Party of China and offers five choices. If a student selects the first, second, and fifth options, the corresponding feature vector would be (1, 1, 0, 0, 1). Additionally, this study encompasses three multiple-choice questions: Questions 8, 16 (which examines sources of academic motivation and comprised eight options), and 22 (which pertains to preferences for participation in activities, with eight options). The detailed feature-handling process for the multiple-choice questions is presented in Table 6.

We divided the dataset into two subsets: 2400 samples were allocated to the training set to train the classifier, and the remaining 92 samples were used to form the test set. After constructing the BP neural network model, we evaluated it using the test set. The architectural design of the college student group-classification model is illustrated in Fig. 5.

In the BP neural network model, the transformed feature vectors from single and multiple-choice questions are selected as the input for the model, resulting in a 34-dimensional input vector for the neural network. The number of nodes in the hidden layer can be adjusted based on the practical requirements. The number of nodes in the output layer is set to four. In the model's output segment, we define the following classification labels: (1,0,0,0) represents Class 1, (0,1,0,0) represents Class 2, (0,0,1,0) represents Class 3, and (0,0,0,1) represents Class 4. Therefore, the output is a 1×4 vector, with the specific class representations listed in Table 7.

5.1. Model setting and parameter selection

To construct the classification model, we opted for the BP neural network as the foundational framework. This model comprises input, hidden, and output layers [30,31]. Fig. 6 depicts the structural design of the classification model based on the BP neural network. During the model training process, we adjust the hyper-parameters to optimize model performance. Specifically, we employ the gradient descent method [32] to optimize the weights and thresholds of neural network nodes and select the cross-entropy loss function as the optimization objective [33–35]. In the process of tuning hyper-parameters, we focus on key parameters such as the learning rate and number of hidden nodes. First, in terms of choosing the learning rate, we conduct meticulous adjustments. The learning rate governs the magnitude of the weight updates in each iteration, and values that are either high or low may lead to a decline in performance. Through multiple experiments and in-depth analysis, we adjust the learning rate to 0.07, striking an optimal balance between the training speed and accuracy [36–38]. Second, we set the number of hidden nodes, which is a critical parameter that influences the complexity and learning capacity of the neural network model. Through iterative experiments and a comprehensive assessment of model performance, we set the number of hidden nodes to 17 to ensure model effectiveness while avoiding overfitting, and set the overall iteration count of the model to 10,000, with the loss function value outputted every 1000 iterations to monitor the training progress. Based on the training progress, we further adjust the parameters to ensure a continuous decrease in the loss function. During the last iteration, the change in the loss function is approximately 0.01, indicating that the variation is sufficiently small. Throughout the entire training process, we ensure that the model not only converges to the appropriate solution during the learning process but also generalizes to unseen data, preventing overfitting. Ultimately, we conclude that the model has an accuracy of 90.22%, indicating that our model demonstrates outstanding performance in this research task.

5.2. Comparison analysis with other classification models

After conducting performance tests on the model, we explore whether this BP neural network-based classification model outperforms other classification methods in terms of performance. To answer this question, we introduce mainstream multiclass algorithms, including XGBoost [39] and the Naive Bayes classifier [40], for comparison. Considering the stochastic nature of the model training process, we perform ten training runs for each algorithm and calculated the average accuracy. Table 8 presents the final

Table 5
Single-choice question option feature examples.

Number	Question 9	Question 10	…	Question 13	Question 14	Question 15	Category
1	4	5	…	4	4	4	3
2	4	5	…	4	4	3	3
3	5	5	…	5	5	5	4
4	5	5	…	3	3	3	2
5	5	5	…	3	3	3	3
6	5	5	…	4	2	3	3
7	3	5	…	5	3	3	4
8	3	4	…	3	3	3	3
9	5	5	…	4	3	3	3
10	5	5	…	4	4	3	4
11	5	5	…	5	5	5	4
…	…	…	…	…	…	…	…

Table 6
Multiple-choice question option data examples.

Sample	Question 8					Question 16						Question 22							
1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	0	0
2	0	0	1	0	0	1	0	1	1	0	0	0	1	0	1	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0
4	1	1	0	1	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0
5	1	1	1	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0
6	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
7	1	1	1	1	1	0	0	0	0	0	0	1	0	1	1	0	1	0	0
8	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
9	0	1	0	1	0	0	0	0	1	1	0	0	0	0	1	0	1	0	0
10	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	0	1	0	0
11	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0
12	1	1	1	0	0	1	0	1	0	0	0	0	0	1	1	0	1	0	0
13	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
14	0	1	0	0	0	1	1	1	0	1	0	0	0	1	1	1	0	0	0
…	…							…											…

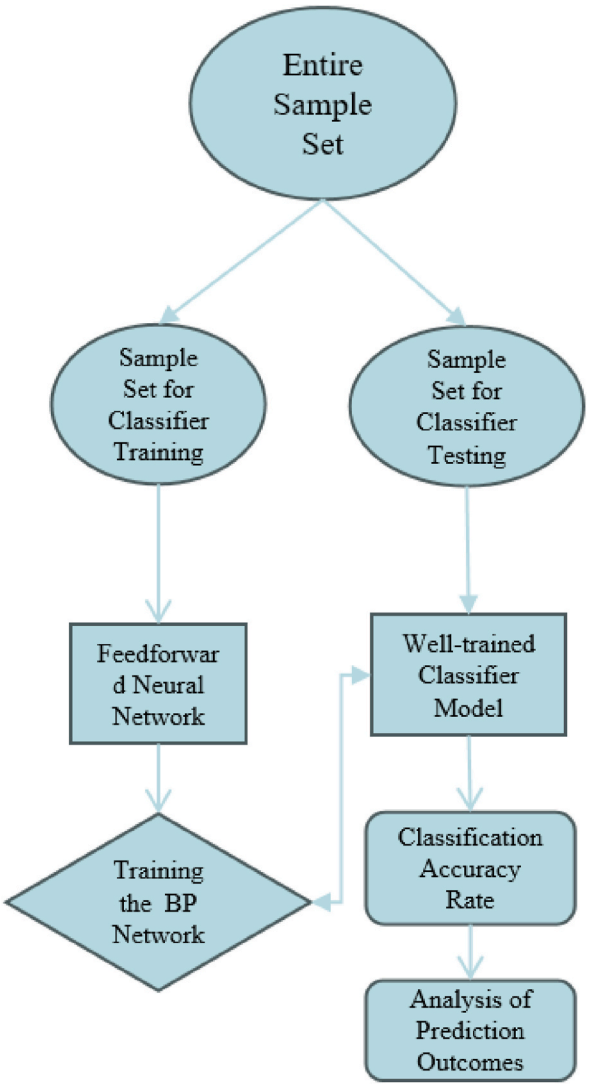


Fig. 5. Overall framework design of the classification model.

Table 7
Classification indicators.

Model Output	Category	Category Name
(1,0,0,0)	1	Diligent Study Type
(0,1,0,0)	2	Pragmatic Type
(0,0,1,0)	3	Discerning Type
(0,0,0,1)	4	Moral Cultivation Type

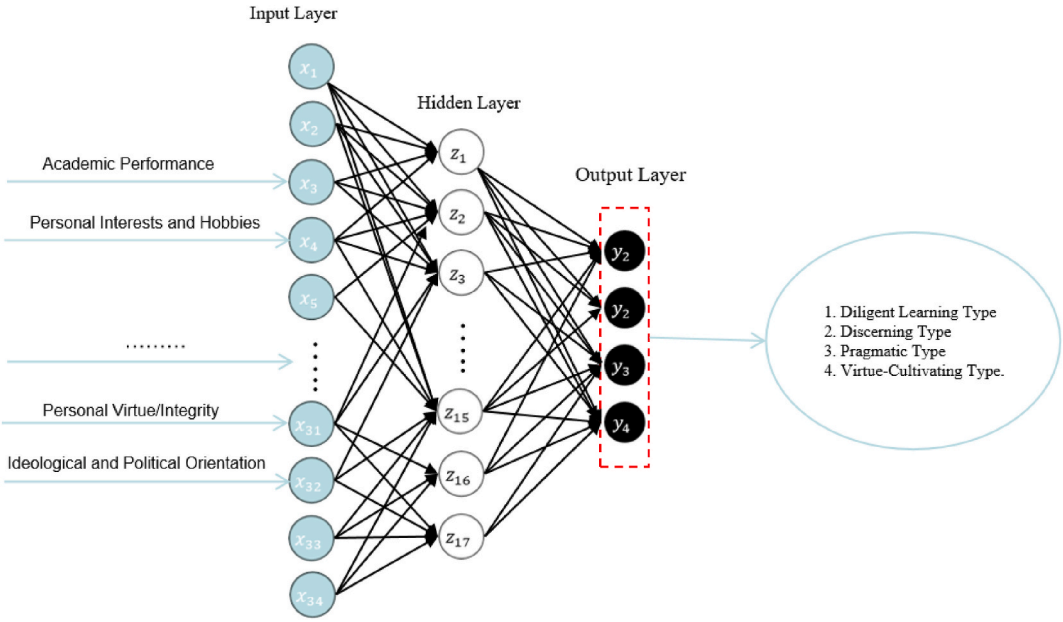


Fig. 6. Structure design of classification model based on BP neural network.

training and comparison results.

6. Conclusion

This study initially utilizes the K-means clustering model to develop detailed and accurate profiles of university student groups. These profiles enable the formation of tailored educational programs and career guidance for distinct student cohorts, significantly contributing to the educational sector. The study integrates the K-means clustering algorithm with the BP neural network model, thereby improving the analysis of survey data. This integration overcomes the limitations inherent in traditional methods that often fail to fully utilize questionnaire data. The methodology demonstrates versatility across various analytical domains, including corporate client credibility assessment, key user segment identification in electronic products, employee satisfaction measurement, and evaluation of primary and secondary students' academic standing, offering fresh perspectives and methodologies for future survey research.

Nonetheless, the study has certain limitations that can be addressed in future research. First, our sample is predominantly drawn from specific colleges or regions, potentially introducing geographical and cultural biases that limit the generalizability of the findings. Second, the reliance on survey data may compromise objectivity and accuracy due to respondent subjectivity and memory biases. Third, K-means clustering and BP neural network models may not fully capture complex data patterns and uncertainties, possibly overlooking subtle but important information. Future research could broaden the sample base to enhance representativeness, diversify data collection methods to reduce biases, and utilize more sophisticated analytical techniques such as deep learning and natural

Table 8
Performance comparison of BP neural network and other classification algorithms.

Classification methods	Average accuracy
BP Neural Network	90%
XGBoost	65%
Naive Bayes Classifier	89%

language processing. Moreover, fostering interdisciplinary collaborations and promoting the application of student profiling technologies in educational management, guidance, and career planning could enhance the impact of this research.

Ethics statement

This study was reviewed and approved by the Ethics Committee of Shaoxing University, with the approval number [YXRQ-2023-001].

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Ran Song: Writing – original draft, Project administration. **Fei Pang:** Investigation, Funding acquisition, Data curation. **Hongyun Jiang:** Validation. **Hancan Zhu:** Writing – review & editing, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Humanities and Social Science Fund of Ministry of Education of China (22JDSZ3107).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e29181.

References

[1] K. Durairaj, I.N. Umar, A proposed conceptual framework in measuring social interaction and knowledge construction level in asynchronous forum among university students, Procedia-Soc. Behav. Sci. 176 (2015) 451–457.
[2] M. De Laat, V. Lally, L. Lipponen, R.-J. Simons, Investigating patterns of interaction in networked learning and computer-supported collaborative learning: a role for social network analysis, Int. J. Comp.-Supported Collab. Learning 2 (2007) 87–103.
[3] F. Colace, M. De Santo, M. Vento, Evaluating on-line learning platforms: a case study, in: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003, 2003, p. 9.
[4] C.N. Gunawardena, C.A. Lowe, T. Anderson, Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing, J. Educ. Comput. Res. 17 (1997) 397–431.
[5] H.-T. Hou, Exploring the behavioral patterns of learners in an educational massively multiple online role-playing game (MMORPG), Comput. Educ. 58 (2012) 1225–1233.
[6] M.H. Jantti, B. Cox, Measuring the value of library resources and student academic performance through relational datasets, Evid. Base Libr. Inf. Pract. 8 (2013) 163–171.
[7] C. Romero, S. Ventura, Educational data mining and learning analytics: an updated survey, Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 10 (2020) e1355.
[8] Á.F. Agudo-Peregrina, S. Iglesias-Pradas, M.Á. Conde-González, Á. Hernández-García, Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning, Comput. Hum. Behav. 31 (2014) 542–550.
[9] O. Iatrellis, A. Kameas, P. Fitsilis, Academic advising systems: a systematic literature review of empirical evidence, Educ. Sci. 7 (2017) 90.
[10] P.H. Wu, G.J. Hwang, M. Milrad, H.R. Ke, Y.M. Huang, An innovative concept map approach for improving students' learning performance with an instant feedback mechanism, Br. J. Educ. Technol. 43 (2012) 217–232.
[11] H. Zhao, Y. Zuo, C. Xu, H. Li, What are students thinking and feeling? Understanding them from social data mining, Int. J. Comput. Appl. Technol. 65 (2021) 110–117.
[12] M. Constantinides, J. Dowell, A framework for interaction-driven user modeling of mobile news reading behaviour, in: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, 2018, pp. 33–41.
[13] R. Asif, A. Merceron, S.A. Ali, N.G. Haider, Analyzing undergraduate students' performance using educational data mining, Comput. Educ. 113 (2017) 177–194.
[14] Q. Yuan, G. Cong, K. Zhao, Z. Ma, A. Sun, Who, where, when, and what: a nonparametric bayesian approach to context-aware recommendation and search for twitter users, ACM Trans. Inf. Syst. 33 (2015) 1–33.
[15] A. Shen, R. Tong, Y. Deng, Application of classification models on credit card fraud detection, in: 2007 International Conference on Service Systems and Service Management, 2007, pp. 1–4.
[16] S. Thomassey, A. Fiordaliso, A hybrid sales forecasting system based on clustering and decision trees, Decis. Support Syst. 42 (2006) 408–421.
[17] F. Angiulli, Fast nearest neighbor condensation for large data sets classification, IEEE Trans. Knowl. Data Eng. 19 (2007) 1450–1464.
[18] B. Şen, E. Uçar, D. Delen, Predicting and analyzing secondary education placement-test scores: a data mining approach, Expert Syst. Appl. 39 (2012) 9468–9476.
[19] P.H.B. Ruas, A.D. Machado, M.C. Silva, M.R. Meireles, A.M.P. Cardoso, L.E. Zárate, C.N. Nobre, Identification and characterisation of Facebook user profiles considering interaction aspects, Behav. Inf. Technol. 38 (2019) 858–872.
[20] S. Mulder, Z. Yaar, The User Is Always Right: A Practical Guide to Creating and Using Personas for the Web: New Riders, 2006.

[21] M.R. Mishra, N. Kaushik, Customer segmentation in mobile services industry A cluster and VALS 2 systems approach, in: *Discovery Summit-2012*, SAS World Headquarters, NC, USA, 2012.
[22] P.S. Bradley, K.P. Bennett, A. Demiriz, Constrained k-means clustering, Microsoft Res., Redmond 20 (2000).
[23] L. Ma, M.M. Crawford, J. Tian, Local manifold learning-based $ k $-nearest-neighbor for hyperspectral image classification, IEEE Trans. Geosci. Rem. Sens. 48 (2010) 4099–4109.
[24] D.T. Pham, S.S. Dimov, C.D. Nguyen, Selection of K in K-means clustering, Proc. IME C J. Mech. Eng. Sci. 219 (2005) 103–119.
[25] M. Ahmed, R. Seraj, S.M.S. Islam, The k-means algorithm: a comprehensive survey and performance evaluation, Electronics 9 (2020) 1295.
[26] K.P. Sinaga, M.-S. Yang, Unsupervised K-means clustering algorithm, IEEE Access 8 (2020) 80716–80727.
[27] S. Ding, C. Su, J. Yu, An optimizing BP neural network algorithm based on genetic algorithm, Artif. Intell. Rev. 36 (2011) 153–162.
[28] P.J. Werbos, Backpropagation through time: what it does and how to do it, Proc. IEEE 78 (1990) 1550–1560.
[29] J. Shore, R. Johnson, Properties of cross-entropy minimization, IEEE Trans. Inf. Theor. 27 (1981) 472–482.
[30] H. Chockler, E. Farchi, B. Godlin, S. Novikov, Cross-entropy based testing, in: Formal Methods in Computer Aided Design (FMCAD'07), 2007, pp. 101–108.
[31] J. Zhang, Gradient Descent Based Optimization Algorithms for Deep Learning Models Training, 2019 *arXiv preprint arXiv:1903.03614*.
[32] P.-T. De Boer, D.P. Kroese, S. Mannor, R.Y. Rubinstein, A tutorial on the cross-entropy method, Ann. Oper. Res. 134 (2005) 19–67.
[33] A. Jamin, A. Humeau-Heurtier, (Multiscale) cross-entropy methods: a review, Entropy 22 (2019) 45.
[34] W. Jin, Z.J. Li, L.S. Wei, H. Zhen, The improvements of BP neural network learning algorithm, in: WCC 2000-ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000, 2000, pp. 1647–1649.
[35] S.E. Buttrey, Data mining algorithms explained using R, J. Stat. Software 66 (2015) 1–4.
[36] M.I. Jordan, T.M. Mitchell, Machine learning: trends, perspectives, and prospects, Science 349 (2015) 255–260.
[37] Q. Bi, K.E. Goodman, J. Kaminsky, J. Lessler, What is machine learning? A primer for the epidemiologist, Am. J. Epidemiol. 188 (2019) 2222–2239.
[38] X. Ying, An overview of overfitting and its solutions, in: Journal of Physics: Conference Series, 2019 022022.
[39] A. Ogunleye, Q.-G. Wang, XGBoost model for chronic kidney disease diagnosis, IEEE ACM Trans. Comput. Biol. Bioinf 17 (2019) 2131–2140.
[40] S. Mukherjee, N. Sharma, Intrusion detection using naive Bayes classifier with feature reduction, Procedia Technol. 4 (2012) 119–128.