J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 8829--8844

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on Prediction and Model Construction of Innovative and Entrepreneurial Mental Health Status Based on Random Forest Algorithm

Li Huang[1, ✉]

[1] Hunan High-speed Railway Vocational and Technical College, Hengyang, Hunan, 421200, China

ABSTRACT

The study constructs a prediction model to predict the mental health status of innovative entrepreneurs. The real data of mental health assessment of innovative entrepreneurs in S province in 2023 is chosen as the data source. The recursive random forest feature elimination method is used to select the features of the mental health status prediction model. The pre-selection-elimination mechanism was used to construct the mental health state prediction model. The prediction models constructed by support vector machine algorithm, decision tree algorithm and random forest algorithm were trained and evaluated respectively. The AUC value and accuracy corresponding to the random forest algorithm are 0.9126 and 86.39%, respectively, which are better than the other two comparison models. Among the 17 mental health characteristic variables selected in this paper, emotional stress and self-acceptance degree have the greatest influence on the prediction model based on the random forest algorithm.

Keywords: Support vector machine, Decision tree, Random forest, Mental health status

1. Introduction

In modern society, innovative and entrepreneurial ability is regarded as an important element in the future career development of college students. However, the mental health problems of college students have gradually attracted widespread attention. Therefore, understanding the relationship between college students' mental health and innovation and entrepreneurship ability is crucial to help college students develop healthily [1-3].

First of all, mental health has an important influence on the development of college students' innovation and entrepreneurship ability. Mental health refers to the good state of an individual in terms of mental, emotional and social functions. Under good psychological state, college students are

✉ Corresponding author.
E-mail address: 15073405777@163.com (L. Huang).
Received 20 February 2024; Revised 15 April 2024; Accepted 20 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-490

more likely to have positive innovation and entrepreneurship consciousness and behavior [4-6]. On the contrary, if college students have mental health problems, such as anxiety, depression, or excessive stress, it may affect their thinking ability, emotional regulation and decision-making ability, which may reduce the innovation and entrepreneurship ability [7-9]. Secondly, innovation and entrepreneurship ability also has a positive effect on the enhancement of college students' mental health. The success and sense of achievement in the process of innovation and entrepreneurship will enhance the self-esteem and self-confidence of college students, thus improving their psychological health [10-12]. In addition, the difficulties and challenges in the process of innovation and entrepreneurship can also motivate college students to face difficulties positively, enhance their ability to cope with pressure, and thus improve their mental toughness and resilience [13-14]. However, the relationship between college students' mental health and innovation and entrepreneurship ability is interactive. Mental health problems may become obstacles to the development of college students' innovation and entrepreneurship ability, but at the same time, they may also be improved and promoted from the process of innovation and entrepreneurship [15-17].

Literature [18] reveals the influence of college students' psychological quality on their innovation and entrepreneurship. The integration of students' mental health education and innovation and entrepreneurship is examined in the light of the educational practice of college students. It is emphasized that educational concepts should keep pace with the times, and the combination of the two is beneficial to the development of education. Literature [19] emphasized the importance of college students' psychological quality on entrepreneurship. It analyzes the penetration of mental health education in the cultivation of college students' entrepreneurship education, which is of great practical significance. Literature [20] is based on the development of innovation and entrepreneurship education to alleviate the employment anxiety of college students. The cross-sectional questionnaire method was used to discuss the impact of the development of innovation and entrepreneurship education on college students' employment anxiety. The results showed that innovation and entrepreneurship education intervention had a significant mitigating effect on college students' employment anxiety. Literature [21] mentioned the strategy of integrating red culture and ideological and political education based on educational psychology. The effectiveness of the strategy was proved through comparative experiments. Literature [22] reveals the status quo of many mental health problems of college students, and proposes a teaching mode that combines mental health education with ideological and political education in colleges and universities as a means of intervening in students' psychological activities, and the results prove that the method is effective in improving the mental health problems of college students, and improves the effect of teaching. Literature [23] aims to reveal the cultivation of innovation and entrepreneurship of college students and its impact on talent cultivation and cultural diversity. Its findings provide a scientific theoretical basis for higher education to carry out innovation and entrepreneurship education and cultural and nutritional activities, and have certain practical value. Literature [24] discusses the impact of health on entrepreneurial behavior and the creation of value in entrepreneurial behavior, aiming to stimulate relevant research, i.e., to stimulate the attention of academic practitioners to this social issue. Literature [25] realizes the transformation of the innovation and entrepreneurship education system based on educational psychology, and examines its impact on the positive psychology of college students' entrepreneurship, so as to help college students understand the market entrepreneurial environment and their own entrepreneurial ability, and maintain a positive mindset in entrepreneurial practice. Literature [26] conducted a questionnaire survey on college students. The factors affecting college students' sense of responsibility for innovation and entrepreneurship were analyzed using the factor analysis model, and the relationship between positive qualities and innovation and entrepreneurship was analyzed using linear regression. The results specified that psychological qualities such as modesty, bravery, and wisdom are the factors affecting college students' sense of responsibility for innovation and entrepreneurship. And wisdom is the most obvious predictor of social responsibility and group responsibility. Literature [27] analyzes the relationship between innovation and entrepreneurship education and entrepreneurial intention based on the perspective of educational psychology. A questionnaire survey was conducted on college students, and multiple linear regression analysis and mediation effect model analysis were adopted. The results emphasize that enterprise environment education is positively correlated with entrepreneurial intention. And entrepreneurial knowledge plays a mediating role in the relationship between entrepreneurial intention and entrepreneurial performance.

The study firstly collects the mental health data of innovative entrepreneurs, and adopts data cleaning, data integration, and data transformation to deal with abnormal data. Second, the features that are highly correlated with the mental health status are screened out by recursive random forest feature elimination method. Then, the AUC value, accuracy, precision, recall, F1 value, and loss function value of each model were compared respectively, and the pre-selection-elimination mechanism was used to screen out the best prediction model for mental health state prediction from the support vector machine, decision tree, and random forest models. Finally, SHAP values were used to interpret the models and the contribution of each feature variable to the model output was calculated.

2. Analysis of mental health assessment data

2.1. Data processing

The data used in this paper come from the results of the mental health assessment of innovative entrepreneurs in Province S in 2023, which utilized the symptom self-assessment scale SCL-90.A total of 796 SCL-90 scales were recovered using the mental health assessment system in this paper, of which 750 were valid. The assessment data mainly include the basic information of the innovative entrepreneurs (including gender, household type, family economy, whether they are only child, age, etc.), as well as the scores of the 90 assessment scales.

There are some incomplete and invalid noise data in the data used, and the direct use of the data for experiments will have a certain impact on the experimental results, and the complex data will also have an impact on the efficiency of the implementation of the algorithm to a certain extent. In order to ensure that the ideal experimental results, this paper carries out the following pre-processing operations on the data.

1) Data screening: determine the target of data mining and keep only the data related to the target. For this study, it is necessary to screen the data related to this goal from the massive amount of raw data. Psychological assessment data include information such as the user's anxiety dimension assessment score, depression dimension score, obsessive-compulsive disorder dimension score, etc., as well as basic information such as the assessment user's gender, hukou type, family economy, whether he or she is an only child, and his or her age.

2) Data Cleaning: Including missing fill, error detection, duplicate filtering, consistency test, etc. Missing padding includes zero-value padding, mean-value padding and probability distribution padding. Error detection refers to checking whether the data meets the specifications through the rules. Duplicate filtering refers to duplicate data generated during data export or system failure, which needs to be eliminated or merged to ensure data uniqueness. Consistency testing includes whether the basic information of the innovative entrepreneurs and the information of the psychological assessment data of the innovative entrepreneurs are consistent, and the inconsistent data should be corrected by rules or handled manually.

3) Data conversion: convert the cleaned data into a data format that facilitates data analysis and mining. One conversion method is to convert the time format of the psychometric scale assessment time, assessment time spent, etc. into a unified format. Another is to convert all Chinese characters in the assessment database with English acronyms to reduce the data storage space, for example, converting depressive symptoms into YY, anxiety into JL, and so on.

2.2. Analysis of basic user attributes

The statistical analysis of basic user information data is shown in Table 1. In the sample of the evaluators, most of the innovative entrepreneurs are male (69.7%), have better family financial conditions (56.1%), and are between 20 and 30 years old (64.4%).

Table 1. User basic information data analysis

Variable		Counting	Percentage
Gender	Male	523	69.7%
	Female	227	30.3%
Growth environment	City	498	66.4%
	Countryside	252	33.6%
Only child	Yes	445	59.3%
	No	305	40.7%
Family economy	Worse	106	14.1%
	Medium	223	29.7%
	Better	421	56.1%
Age distribution	20~25	269	35.9%
	26~30	214	28.5%
	31~40	161	21.5%
	>40	106	14.1%

2.3. Analysis of SCL-90 scale data

The dataset for this paper is derived from the Symptom Self-Rating Scale SCL-90, which is globally recognized as one of the most prominent mental health assessment scales and is currently the most widely used screening tool for mental disorders and illnesses in outpatient populations. The SCL-90 examines ten dimensions to assess a person's mental health [28], which consists of 90 items assessing the state of mental health of the innovator-entrepreneurs in different domains of mental health. These include somatization, obsessive-compulsive symptoms, interpersonal sensitivity, depression, anxiety, hostility, fear, paranoia, and psychoticism.Each item on the SCL-90 scale is rated on a five-point scale ranging from 1 (none) to 5 (severe), indicating the severity of the symptoms experienced. The final results include total and mean scores for each dimension, as well as the number of positive items and an overall assessment of the number of positive items, which represents the number of items with scores between 2 and 5, indicating the presence of symptoms described in those specific items.

The SCL-90 scale measure factor score statistics are shown in Figure 1. A total of 413 innovative entrepreneurs had a scale score of more than 180, and their positive detection rate was 55.1%. Among them, the innovative entrepreneurs who showed symptoms of anxiety, symptoms related to interpersonal sensitivity, and symptoms of depression accounted for 49.3%, 26.5%, and 23.1% of the total number of mental health abnormalities, respectively. Among the different symptoms, anxiety seems to be the most prevalent phenomenon among innovative entrepreneurs.

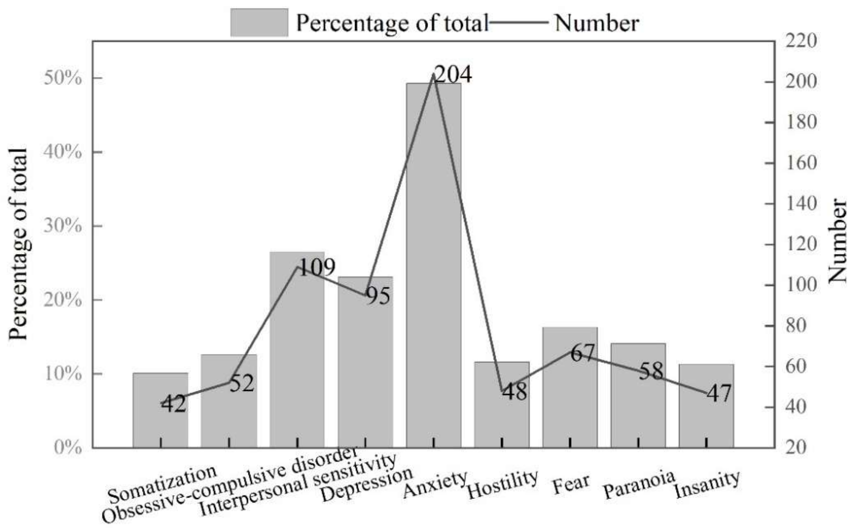


Fig. 1. SCL-90 scale measurement factor score statistics

3. Characterization of models for predicting mental health status

Recursive Feature Elimination (RFE) is an efficient and practical feature selection algorithm, whose core idea is based on a recursive strategy to finely rank features by continuously evaluating their importance. In this process, the algorithm will gradually eliminate those features that contribute relatively little to the model, aiming to achieve model simplification and optimization. This approach not only helps to reduce the complexity of the model, but also significantly improves the generalization ability of the model, which makes the model show stronger robustness and prediction accuracy in the face of new data. The advantage of recursive feature elimination is that it systematically identifies and retains the features that are most critical to model prediction, while removing redundant or noisy features, thus providing a powerful tool for building more streamlined and efficient machine learning models.

RFE relies on the base model to compute the importance scores of features. In the case of a random forest, for example, the importance score of a feature is calculated by relying mainly on the sum of the Gini impurity reduction of that feature in all trees. For a Gini impurity at node t the formula is as follows:

$$G(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

where p_i is the proportion of samples belonging to the category. C is the total number of categories.

The importance of feature j is calculated by accumulating the reduction in Gini impurity contributed by that feature in all trees, with the following formula:

$$\mathrm{Im}\,por\tan ce(j) = \sum_{t \in T} \Delta G(t) \cdot I_t(j) \tag{2}$$

where $\Delta G(t)$ is the Gini impurity reduction at node t. $I_t(j)$ is whether feature j is used for splitting at node t, and T is the set of nodes of all trees.

The specific algorithm is as follows:

Input:

1) Feature matrix X (of shape $n \times m$, where n is the number of samples and m is the number of features).

2) Target vector y (length n).

3) Base model (random forest).

4) Number of features to be selected k.

Output:

The optimal subset of features to be selected:

1) Initialization: set the current feature set to all features $X = \{x_1, x_2, \ldots, x_m\}$. Set the number of features to be selected k.

2) Train the base model using the current feature set X.

3) Calculate the importance score of each feature according to equations 1) and 2).

4) Sort the features according to the importance score.

5) Remove the feature with the lowest importance score.

6) Repeat steps 2) to 5) until the number of remaining features is equal to k.

7) Output the optimal subset of features and return the current feature set X.

The advantage of RFE is that it can significantly reduce the number of features, which reduces the complexity of the model and improves the generalization ability of the model. At the same time, the interpretability of the model can be enhanced by eliminating irrelevant or redundant features.

When using RFE for feature selection, there is a certain degree of blindness in setting n_features_to_select, which may make the model performance deteriorate, so the Recursive Random Forest Feature Elimination Method (RF-RFECV), which is combined with cross-validation, is used in practice to compute the validation errors of all the subsets, and ultimately select the features selected from the subset with the smallest error rate.

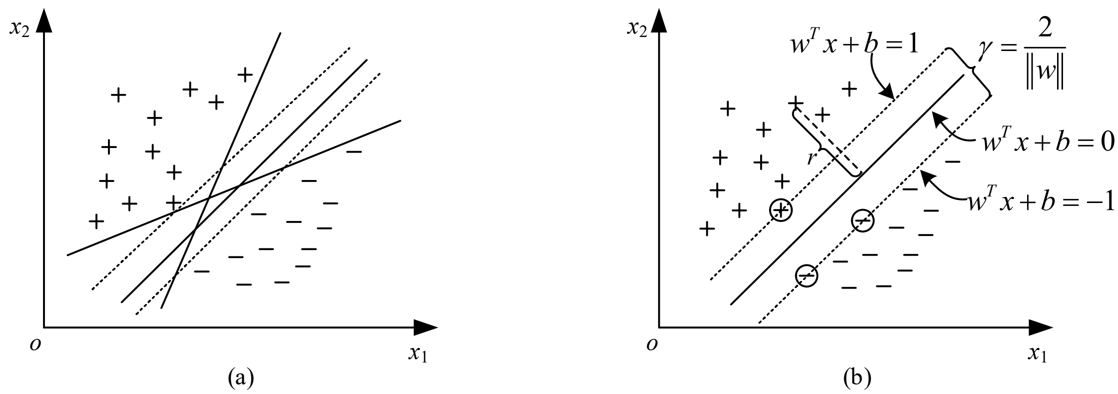
4. Construction of a model for predicting mental health status

The study uses a pre-selection-elimination mechanism to construct a mental health state prediction model. Pre-selection means that several different types of classification models with classification ability are first selected, given the same training dataset for training, and finally, based on the average results of several experiments, the model with the optimal output is selected as the final prediction model, and the rest of the models are eliminated. The reason for choosing this method to construct the model is that the mental health assessment data used in this paper is a kind of data that has not been studied and applied, so it is necessary to pre-select some models to compare and analyze the advantages and disadvantages of the performance of the different models for this kind of data, and then finally determine the prediction model.

4.1. Modeling algorithm selection

4.1.1. Support vector machine algorithms. The essence of Support Vector Machine (SVM) is to divide the sample data identified as different classes in the feature dimension space with at least one hyperplane, which is also the classification boundary between these classes can also be called as decision boundary [29].

The division of the sample data classes is shown in Figure 2. When the sample data is linearly divisible as shown in Fig. 2(a), it can be seen that there can be multiple straight lines or hyperplanes between the positive and negative classes to distinguish them from each other. However, the purpose of the support vector machine is to find a straight line or hyperplane so that the two classes can be distinguished as much as possible as shown in Fig. 2(b), and for this purpose the SVM uses the interval maximization method to find this straight line or hyperplane.



(a) Linear Separable Data Set (b) Maximum Interval Hyperplane

Fig. 2. Sample size is classified according to category

The dataset with m sample is $D=\{(X_1,y_1),(X_2,y_2),\cdots,(X_m,y_m)\}$, y_m denotes the classification identifier of the m rd data sample, the feature space of the sample dataset contains n feature dimensions, and the feature dimension vector corresponding to the m th sample data is $X_m=\{x_m^1,x_m^2,\cdots,x_m^n\}$.

A linearly differentiable hyperplane H can be represented as:

$$H:\omega^T\cdot X+b=0 \tag{3}$$

where $\omega^T=\{\omega_1,\omega_2,\cdots,\omega_n\}$ is the weight vector and b denotes the bias. Then the classification boundary of this hyperplane is:

$$H_1:\omega^T X+b\geq 1, y=+1 \tag{4}$$

$$H_{-1}:\omega^T X+b\leq -1, y=-1 \tag{5}$$

The interval between hyperplane classification boundaries can be expressed as:

$$\gamma=\frac{2}{\|\omega\|} s.t. y_i\left(\omega^T x_i+b\geq 1\right), i=1,2,\cdots,n \tag{6}$$

Maximizing γ under constraints is the objective of the support vector machine to maximize the interclass spacing, and the objective function is:

$$f=\min_{\omega,b}\frac{1}{2}\|\omega\|^2 \ s.t. y_i\left(\omega^T x_i+b\geq 1\right), i=1,2,\cdots,n \tag{7}$$

The training procedure for the linearly divisible SVM algorithm is shown below:

Input: linearly differentiable training set D.

Output: hyperplane and classification model.

1) Solve the optimization problem with constraints by combining the KKT conditions:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i\left(x_i\cdot x_j\right)-\sum_{i=1}^{N}\alpha_i \tag{8}$$

$$s.t.\sum_{N}^{i=1}\alpha_i y_i=0, \alpha_i\geq 0, i=1,2,\cdots,N \tag{9}$$

Find the optimal solution: $\alpha^* = \left(\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*\right)^T$.

2) Calculate:

$$\omega^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i \tag{10}$$

Choose a positive component $\alpha_j^* > 0$ of α^* and compute:

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i \left(x_i \cdot x_j\right) \tag{11}$$

3) Find the separating hyperplane:

$$\omega^* \cdot x + b^* = 0 \tag{12}$$

Separate decision functions:

$$f\left(x\right) = sign\left(\omega^* \cdot x + b^*\right) \tag{13}$$

When the sample dataset is linearly divisible, the support vector machine divides the dataset by means of hyperplanes with low dimensions, while when the dataset is nonlinearly divisible, the support vector machine handles it in the opposite way to the operation when it is linearly divisible, and maps the dataset to high dimensions by means of kernel functions, in order to find the existence of hyperplanes that can divide the dataset in a high-dimensional space. The kernel function is defined as:

Let χ be a n-dimensional space, H a m-dimensional space, and $m \gg n$ if there exists a mapping $\varphi(x): \chi \to H$ from n to m dimensions such that for all $x, z \in \chi$ the function $K\left(x, z\right)$ satisfies the following conditions:

$$K\left(x, z\right) = \left\langle \varphi\left(x\right), \varphi\left(z\right) \right\rangle = \varphi\left(x\right) \cdot \varphi\left(z\right) \tag{14}$$

Then function $K\left(x, z\right)$ is the kernel function and $\varphi\left(x\right)$ is the mapping function.

The training procedure of the nonlinearly differentiable SVM algorithm is shown below:

Input: nonlinearly differentiable training set D.

Output: classification model.

1) Select appropriate kernel function $K\left(x, z\right)$ and parameters C to solve the optimization problem under constraints:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K\left(x_i \cdot x_j\right) - \sum_{i=1}^{N} \alpha_i \tag{15}$$

$$s.t. \sum_{N}^{i=1} \alpha_i y_i = 0, 0 \le \alpha_i \le C, i = 1, 2, \cdots, N \tag{16}$$

Find the optimal solution:

$$\alpha^* = \left(\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*\right)^T \tag{17}$$

2) Choose a positive component $0 < \alpha_j^* < C$ of α^* and compute:

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i K\left(x_i \cdot x_j\right) \tag{18}$$

3) Decision function:

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x \cdot x_i) + b^*\right) \tag{19}$$

4.1.2. Decision Tree Algorithm. The decision-making mechanism of decision tree is a learning method that makes constant judgment according to the characteristic attributes of the data set and finally achieves the classification effect. The model structure of decision tree is tree-shaped and intuitively visible, so the model is highly interpretable. Decision tree through the multi-layer conditions of continuous decision-making, and finally determine the input belongs to the classification type [30]. A decision tree is a tree structure consisting of nodes and directed edges, except for the leaf nodes which represent classifications, the rest of the nodes represent features or attributes and the directed edges represent different decision conditions. There are many decision tree algorithms, of which C4.5 algorithm is described in this section. To understand the C4.5 algorithm one must first understand the following concepts:

Information entropy: a parameter used to measure the orderliness of the data in a data system, when there is a variety of data present within a data system, and the proportion of data from various categories is not low, then the sample set is disordered and impure. The formula for information entropy is:

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \tag{20}$$

D is the dataset, p_k is the proportion of data in category k, and $|y|$ indicates the number of categories; the smaller the information entropy, the purer the sample dataset. The minimum value of 0 indicates that there is only one category of data in the sample dataset.

Information gain: the difference between the information entropy of the sample dataset and the original dataset after it is divided according to a certain feature dimension. The formula for information gain is as follows:

$$Gain(D,\alpha) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{21}$$

D is the dataset, α is the selected attribute, there are V values in α, with these V values to divide the dataset D, respectively, to get the dataset D^1 to D^V, respectively, the entropy of these V datasets, and weighted average, and finally the difference with the information entropy of the dataset D, that is, the information gain.

Information gain rate: the ratio of information gain $Gain(D,\alpha)$ to the information entropy $IV(\alpha)$ of attribute α:

$$Gain\_ratio(D,\alpha) = \frac{Gain(D,\alpha)}{IV(\alpha)} \tag{22}$$

$$IV(\alpha) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|} \tag{23}$$

The decision basis of the C4.5 algorithm is based on the information gain rate of the dataset after each split, and each time the variables are split the direction with the maximum gain rate feedback is selected for execution, and finally the whole decision tree is obtained.

4.1.3. Random Forest Algorithm. Random Forest algorithm is essentially composed of many subtrees, so it is figuratively called "forest", these subtrees are the decision trees mentioned earlier, the decision results of the subtrees will be combined with the Boosting strategy through integrated learning to determine the output of the Random Forest model. The Random Forest algorithm is highly resistant to data interference and has excellent applicability in classification scenarios [31]. The algorithmic process of random forest is as follows:

Input: training dataset D of size N, data with feature dimension M.

Output: random forest model $M = \{T_1, T_2, \cdots, T_K\}$.

1) Sample N data samples D_i from the dataset with putback using bootstrap sampling method.

2) Randomly select $m \square LM$ feature subsets from the feature dimension space.

3) Using the decision tree algorithm, train a decision tree T_i with each tree left to grow without pruning operations.

4) Repeat steps 1) to 3) K times to obtain a random forest model $M = \{T_1, T_2, \cdots, T_K\}$.

4.2. Model Training and Evaluation

Three modeling algorithms were used in the pre-selection modeling algorithm phase, which are Decision Tree Algorithm, Support Vector Machine Algorithm (SVM), and Random Forest Algorithm (RF) to build the classification model. The definition domain of the model and the representation of the function are as follows:

$$U_{u\times b} \times R_{b\times 1} \to R_{u\times 1} \tag{24}$$

U is a matrix of mental health characteristics of a sample of labeled individuals, P is a matrix of mental health states of a sample of labeled individuals, and R is a projection matrix that reveals potential mapping relationships between mental health characteristics and mental health states of labeled individuals. Each user's mental health profile is a b-dimensional feature vector defined as the number of items under a factor of the SCL-90. If we are able to collect the mental health characteristics of a sample of labeled individuals, we can build U. If we are able to collect the Symptom Self-Rating Scale SCL-90 measurements of a labeled sample, we can build P. When both U and P are established, the key R that can predict mental health status in a mental health prediction model can be established. To be able to obtain the optimal R, we define the following object function:

$$f(u,r) = F(u,r) - P_0 \tag{25}$$

P_0 is the test result of the Symptom Self-Rating Scale SCL-90, r is the projection matrix, and the task of this paper is to find a r that minimizes f:

$$f_r = \arg\min f(u,r) \tag{26}$$

The evaluation metrics corresponding to this two-class classification model are set as precision and recall.

The confusion matrix for the two-class classification model is shown in Table 2. Four values are used to calculate the precision and recall which are true positive tp, false positive fp, true negative tn and false negative fn, the sum of these four values is the total number of samples in the sample set i.e. $tp + fp + tn + fn = n$, n is the total number of samples.

Table 2. Confusion matrix

Confusion Marrix		Predict	
		Positive	Negative
Real	Positive	tp	fn
	Negative	fp	tn

The precision rate measures the accuracy of the model's prediction results, and for the two-class classification model, which is divided into positive and negative classes, their precision rates can be calculated separately with the following formula:

$$\text{Precision rate of the positive category} = tp/(tp + fp) \tag{27}$$

$$\text{Precision rate of the negative class} = tn/(tn + fn) \tag{28}$$

Recall measures the rate at which samples in a sample set are successfully predicted, and the formula for recall for the POSITIVE and NEGATIVE classes is as follows:

$$\text{Recall of the positive class} = tp/(tp + fn) \tag{29}$$

$$\text{Recall of the negative class} = tn/(tn + fp) \tag{30}$$

$$\text{Accuracy of the model} = (tp + tn)/n \tag{31}$$

5. Analysis of model experiment results

5.1. Selection of the number of features

Mental health related data collected for this study included a total of 37 variables including general demographic indicators, sleep, exercise, personality and stress. Features were selected using recursive random forests and subsequent predictive models were learned based on the final selection. In order to find the optimal number of features, cross-validation is used along with recursive random forest to obtain accuracy scores for different subsets of features and select the set of features with the best scores. During the process, the number of cross-validation was set to 5 and finally 17 features were selected as a result of RFECV feature selection. The relationship between the number of features and the 50% discount cross-validation score is shown in Figure 3. With the increase of the number of features, the score is in an overall upward trend. When the number of features is 17, the curve appears to peak at 0.8076, while after that the score has a fluctuation as the number of features increases, but the effect is all lower than when the number of features is 17. The 17 characteristics thus selected were: emotional stress, self-acceptance, life feelings, neuroticism, sleep quality, self-evaluation, interpersonal stress, relationship stress, sense of meaning-seeking, life attitudes, life goals, agreeableness, extraversion, health stress, family stress, openness, and frustration stress. The variables obtained after feature selection were all continuous type variables.

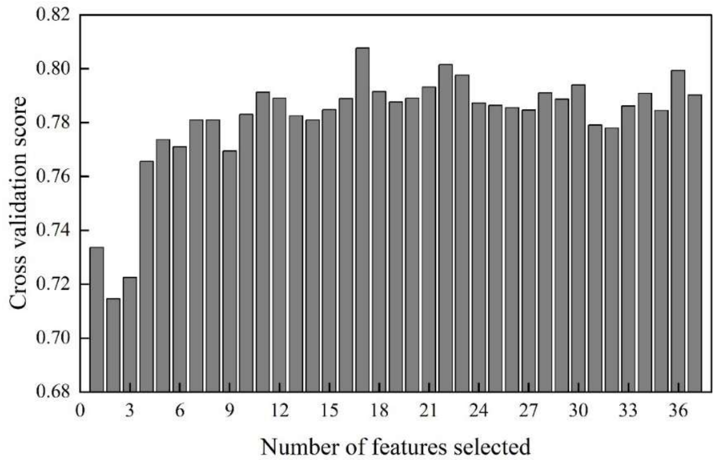


Fig. 3. Feature selection validation

5.2. Analysis of model prediction results

Decision-Tree, SVM, and RF models were utilized for the prediction of mental health status of innovative entrepreneurs, respectively. The models were trained using the training set and the trained models were validated using the validation set. The ROC curves of each model in the test set are shown in Figure 4. The corresponding AUC values for Decision-Tree, SVM, and RF are 0.9032, 0.9049, and 0.9126, respectively.

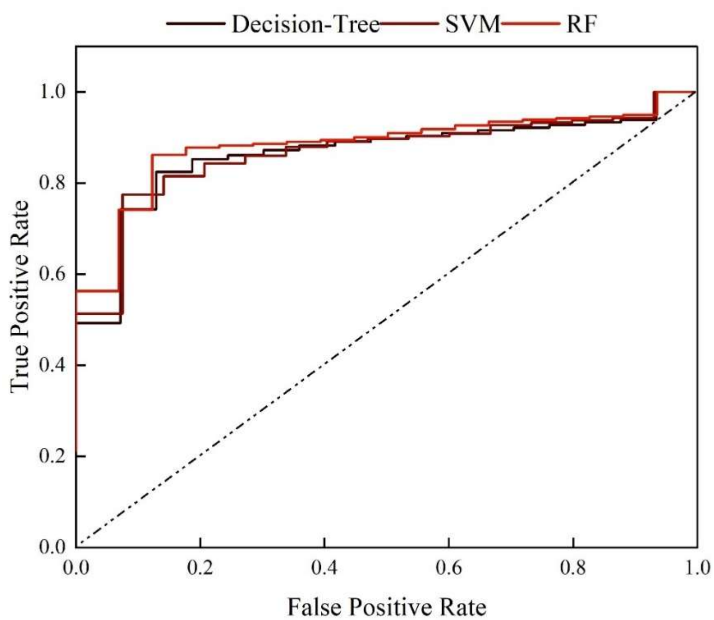


Fig. 4. ROC curve of each model in the test set

The comparison of the accuracy, precision, recall and F1 value of each model is shown in Table 3. The RF model proposed in this paper achieved optimal results in all four evaluation indicators, and its accuracy rate of predicting the mental health status of innovative entrepreneurs was 86.39%.

Table 3. Experimental results of each model

Model	Accuracy/%	Precision/%	Recall/%	F1
Decision-Tree	82.11	65.39	70.36	0.73
SVM	83.64	76.36	70.69	0.74
RF	86.39	89.63	70.98	0.79

The changes in the loss function value of each model with the increase in the number of iterations are shown in Fig. 5. During the model training process the loss function value of the RF model proposed in this paper has been kept lowest compared to the other models, with an average loss function value of 0.194. It shows that the model has strong learning ability and generalization ability, and it can converge quickly and maintains better results.

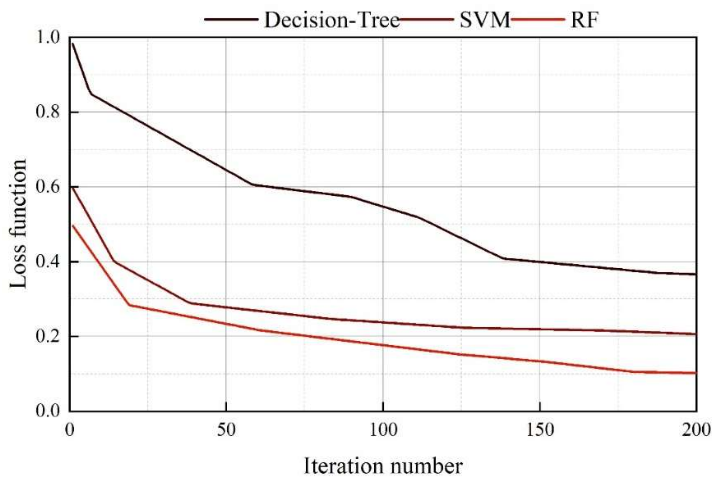


Fig. 5. Changes in the value of loss functions with iteration times

According to the performance of the three models it can be seen that the RF prediction model can get more ideal results. Therefore, the study chooses this Random Forest-based algorithm to construct a prediction model for the mental health status of innovative entrepreneurs, eliminating the Decision-Tree and SVM prediction models.

5.3. Model Interpretation and Characteristic Importance

In order to be able to understand the importance of each variable in the RF model, this study adopts the SHAP value to explain the model.The SHAP value method is an additive explanatory model based on the Shapley value, which adopts a game theoretic approach to explain the output of the model, and it belongs to the ex post explanation method of the model, which outputs a predictive value for each sample, and it can be calculated that the contribution of each variable to the output of the model. Both the influence of a single variable and the influence of a group of variables can be considered, and the problem of multicollinearity can also be solved.The absolute value of the size of the SHAP value represents the degree of influence of the feature, and the sign represents the direction of the influence of the feature. The order of importance of the variables is shown in Figure 6. The variable that has the greatest influence on mental health status is emotional stress (1), followed by self-acceptance (2), and the influence of these two variables in the model is 0.051 and 0.049, respectively.The other variables in descending order of importance are feeling of life (3), neuroticism (4), quality of sleep (5), self-appraisal (6), interpersonal stress (7), and relationship stress (8), Sense of seeking meaning (9), Life attitude (10), Life goals (11), Likability (12), Extraversion (13), Health stress (14), Family stress (15), Openness (16), Frustration stress (17).

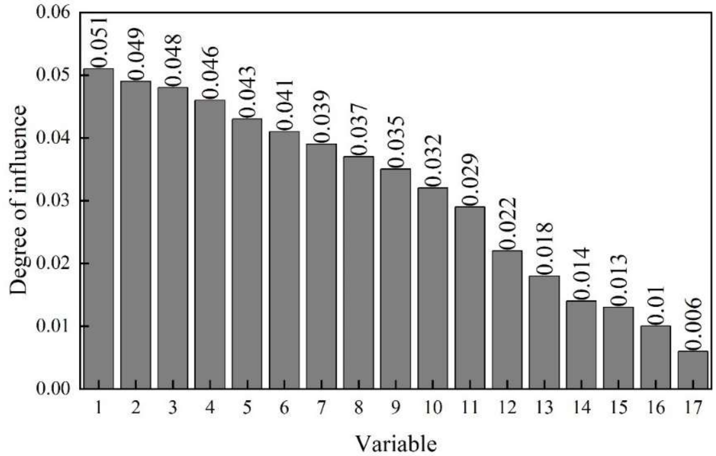


Fig. 6. The importance of each variable sort

6. Conclusion

Support vector machine algorithm, decision tree algorithm and random forest algorithm were chosen to construct the mental health state prediction model. The pre-selection-elimination mechanism was used to select the prediction model with the most ideal prediction effect. The AUC value, accuracy, precision, recall, F1 value, and loss function value corresponding to the random forest algorithm were 0.9126, 86.39%, 89.63%, 70.98%, 0.79, and 0.194, respectively.Compared with the other two models, the prediction model based on the random forest algorithm was more effective in predicting the mental health status of innovative entrepreneurs. Among the 17 feature variables selected using recursive random forest, emotional stress and self-acceptance degree had the greatest degree of influence on the prediction model based on the random forest algorithm.

References

[1] Zhao, L., Li, H., & Chen, L. (2022). Factors influencing Chinese college students' innovation and entrepreneurship ability: The moderating effect test based on entrepreneurial atmosphere. Sustainability, 14(20), 13020.

[2] Aykan, E., Karakuş, G., & Karakoç, H. (2019). The effect of university students' individual innovation and lifelong learning trends on entrepreneurship orientation. Sustainability, 11(22), 6201.

[3] Safin, R. S., Shaidullina, A. R., Alikhanova, R. A., Muskhanova, I. V., Yusupkhadzhieva, T. V., Dzhamalkhanova, L. A., ... & Akhmetov, L. G. (2016). Innovative entrepreneurship in education: a new look in the students training content and existing problems. International Review of Management and Marketing, 6(2), 51-56.

[4] Charles, S. T., Karnaze, M. M., & Leslie, F. M. (2022). Positive factors related to graduate student mental health. Journal of American college health, 70(6), 1858-1866.

[5] Byrd, D. R., & McKinney, K. J. (2012). Individual, interpersonal, and institutional level factors associated with the mental health of college students. Journal of American College Health, 60(3), 185-193.

[6] Iarovici, D. (2014). Mental health issues and the university student. JHU Press.

[7] Castillo, L. G., & Schwartz, S. J. (2013). Introduction to the special issue on college student mental health. Journal of clinical psychology, 69(4), 291-297.

[8] Kitzrow, M. A. (2009). The mental health needs of today's college students: Challenges and recommendations. NASPA journal, 46(4), 646-660.

[9] Auerbach, R. P., Alonso, J., Axinn, W. G., Cuijpers, P., Ebert, D. D., Green, J. G., ... & Bruffaerts, R. (2016). Mental

disorders among college students in the World Health Organization world mental health surveys. Psychological medicine, 46(14), 2955-2970.

[10] Wu, B., & Ye, S. (2022). The influencing factors of innovation and entrepreneurship intention of college students based on AHP. Mathematical Problems in Engineering, 2022(1), 7861631.

[11] Zhu, J., Lei, L., Wu, P., Cheng, B., Yang, X. L., Fu, J., ... & He, F. (2022). The intervention effect of mental health knowledge integrated into ideological and political teaching on college students' employment and entrepreneurship mentality. Frontiers in Psychology, 13, 1002468.

[12] Otchengco Jr, A. M., & Akiate, Y. W. D. (2021). Entrepreneurial intentions on perceived behavioral control and personal attitude: Moderated by structural support. Asia Pacific Journal of Innovation and Entrepreneurship, 15(1), 14-25.

[13] Gao, H., Qiu, Z., Liu, Z., Huang, L., & San, Y. (2016). Research and practice on college students' innovation and entrepreneurship education. In Social Computing: Second International Conference of Young Computer Scientists, Engineers and Educators, ICYCSEE 2016, Harbin, China, August 20-22, 2016, Proceedings, Part II 2 (pp. 36-44). Springer Singapore.

[14] Jiang, Y., Pan, J., Liu, M., & Gao, X. (2022). The relationship between college students' entrepreneurial intention and the teaching quality of innovation and entrepreneurship practice. International Journal of Emerging Technologies in Learning (iJET), 17(12), 61-76.

[15] Huang, H., He, J., Liu, H., Peng, D., & Wei, Z. (2021). THE INFLUENCE OF "INNOVATION AND ENTREPRENEURSHIP TRAINING PROGRAM" ON COLLEGE STUDENT'S INNOVATION QUALITY FROM THE PERSPECTIVE OF PSYCHOLOGY. Psychiatria Danubina, 33(suppl 7), 66-68.

[16] Zhao, N. (2022). RESEARCH ON THE INFLUENCE OF ANXIETY DISORDER ON COLLEGE STUDENTS'INNOVATION AND ENTREPRENEURSHIP. Psychiatria Danubina, 34(suppl 2), 683-683.

[17] Lai, C., & Liu, W. (2022). RESEARCH ON THE EFFECT OF MENTAL HEALTH EDUCATION ON ALLEVIATING COLLEGE STUDENTS'ANXIETY ABOUT INNOVATION AND ENTREPRENEURSHIP FROM THE PERSPECTIVE OF EDUCATIONAL PSYCHOLOGY. Psychiatria Danubina, 34(suppl 2), 250-250.

[18] Wang, B. (2022). Exploration of the path of innovation and entrepreneurship education for college students from the perspective of mental health education. Journal of healthcare engineering, 2022(1), 2659160.

[19] Cao, J. (2022). Exploration on the Cultivation of Undergraduates' Entrepreneurship Ability from the Perspective of Mental Health Education. Journal of Healthcare Engineering, 2022(1), 9668220.

[20] Liang, Q. (2023). Innovation and entrepreneurship education development on employment anxiety of college students. CNS Spectrums, 28(S2), S127-S128.

[21] Bian, Z. (2022). ANALYSIS OF INTERVENTION MEASURES FOR ANXIETY DISORDER OF COLLEGE STUDENTS IN THE PROCESS OF INNOVATION AND ENTREPRENEURSHIP. Psychiatria Danubina, 34(suppl 1), 208-210.

[22] Liu, Y. (2022). RESEARCH ON THE INFLUENCE OF COLLEGE STUDENTS'INNOVATION AND ENTREPRENEURSHIP ON EMPLOYMENT ANXIETY FROM THE PERSPECTIVE OF EDUCATIONAL PSYCHOLOGY. Psychiatria Danubina, 34(suppl 4), 406-406.

[23] Li, M., Wang, T., & Wu, Y. (2021). Impact of innovation and entrepreneurship education in a university under personality psychology education concept on talent training and cultural diversity of new entrepreneurs. Frontiers in Psychology, 12, 696987.

[24] Shepherd, D. A., & Patzelt, H. (2015). The "heart" of entrepreneurship: The impact of entrepreneurial action on health and health on entrepreneurial action. Journal of Business Venturing Insights, 4, 22-29.

[25] Tu, N., & Guo, X. (2022). ANALYSIS ON THE INFLUENCE OF THE TRANSFORMATION OF INNOVATION AND ENTREPRENEURSHIP EDUCATION SYSTEM ON COLLEGE STUDENTS'ENTREPRENEURIAL POSITIVE PSYCHOLOGY. Psychiatria Danubina, 34(suppl 4), 592-592.

[26] Zhang, Q., Liu, C., Wang, Z., & Yang, Z. (2020). The college students' sense of responsibility for innovation and entrepreneurship. Frontiers in Psychology, 11, 2049.

[27] Wang, Q., Sun, Z., & Wu, C. (2021). The impact of university innovation and entrepreneurship education on entrepreneurial intention from the perspective of educational psychology. Frontiers in psychology, 12, 745976.

[28] Xia Chen,Yue Hu,Yuan Deng,Xin Wang,Xiao Yang,Ying Wang... & Huimin Li. (2024). Psychological status of general population 1 year after the outbreak of COVID-19: a cross-sectional study based on SCL-90. Frontiers in Psychology1420834-1420834.

[29] Rong Chen & Yan Zheng. (2024). Research on Power Grid Load Forecasting Method Based on Support Vector Machine. Journal of Physics: Conference Series(1),012004-012004.

[30] Guo Li,Jinfeng Miao,Ping Jing,Guohua Chen,Junhua Mei,Wenzhe Sun... & Suiqiang Zhu. (2024). Development of predictive model for post-stroke depression at discharge based on decision tree algorithm: A multi-center hospital-based cohort study. Journal of psychosomatic research111942.

[31] Mohamed E. Syed,Naqishbandi Tawseef Ahmad,Bukhari Syed Ahmad Chan,Rauf Insha,Sawrikar Vilas & Hussain Arshad. (2023). A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms. Healthcare Analytics.