Problem Chosen	2022 ShuWei Cup Summary Sheet	Team Control Number
C		2022100812763

Summary

Alzheimer's disease (AD) is a progressive neurodegenerative disease with an insidious onset. Therefore, early and accurate diagnosis of Alzheimer's disease and mild cognitive impairment is of great importance.

For problem 1, the baseline data were filtered as the dataset and de-duplicated. The data in the ADNI1 phase are removed because more than half of the data are missing for several data features in the ANDI1 phase. The data features with too many missing rates were removed and the remaining missing data were filled. Subsequently, outliers are discriminated and corrected. Finally, the importance of each indicator was calculated and ranked using random forest to analyze the correlation between data features and Alzheimer's disease.

For problem 2, the data features belonging to structural brain features and cognitive-behavioral features are filtered based on the correlation weights derived from the random forest algorithm in the first question, combined with medical literature and clinical data. Three models of logistic regression, xgboost, and lightgbm are built and cross-validated. Using the voting integration algorithm, we design each classifier weight based on the classification results, confusion matrix, and ROC curve to build an integrated learning classifier as the optimal diagnostic model.

For problem 3, the analysis was performed using the K-means algorithm with the elbow method based on the pre-processed baseline data set. Firstly, the elbow method is used to determine the k-value, and the reliability of the data feature processing is verified by comparing the k-value with the topic requirement into three categories CN, MCI ,AD. The results show that the value of k is taken as 3. Finally, the clusters are clustered by K-means algorithm and visualized by downscaling to two-dimensional planes using PCA. The refinement of MCI into SMC, EMCI, and LMCI is done in a similar way.

For problem 4, the data were first preprocessed and multiple indicators were screened out by stepwise regression analysis model, then the weights were determined and normalized by substituting the entropy weighting method, making a line graph to study the time nodes and the trend of indicator changes, and establishing a time series model to comprehensively analyze the evolutionary pattern, the results showed that LMCI deteriorated faster and more severely than EMC, the evolution of SMC was smoother and would not directly suffer from Demetia, and there was no sign of improvement in AD.

For problem 5, we proposed early intervention and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI, and AD by searching the relevant literature, and wrote a reasonable report by combining the analysis with the appeal model.

Key word: Random Forest,xgboost,K-Means,Time Series,Logistic regression

Content

1. Introduction

1.1 Background

Alzheimer's disease (AD) is a progressive neurodegenerative disease with an insidious onset. It is clinically characterized by a full spectrum of dementia, including memory impairment, aphasia, language impairment, dyscognition, visuospatial skill impairment, executive dysfunction, and personality and behavioral changes, the cause of which remains unknown. It is characterized by a progressive decline in the ability to perform activities of daily living, accompanied by a variety of neuropsychiatric symptoms and behavioral disturbances.

The disease usually proceeds in the elderly, with progressive loss of independent living skills and death from complications within 10 to 20 years after onset. The preclinical stage of Alzheimer's disease, also known as mild cognitive impairment (MCI), is a transitional state between normal and severe. Due to limited knowledge of the disease among patients and their families, 67% of patients are diagnosed as moderate to severe and miss the optimal stage of intervention. Therefore, early and accurate diagnosis of Alzheimer's disease and mild cognitive impairment is of great importance.

1.2 Work

Question 1: Preprocess the characteristic indicators of the attached data to investigate the correlation between data characteristics and the diagnosis of Alzheimer's disease.

Question 2:Use the attached structural brain features and cognitive behavioral features to design an intelligent diagnosis of Alzheimer's disease.

Question 3:First, cluster CN, MCI and AD into three major classes. Then, for the three subclasses contained in MCI (SMC, EMCI, and LMCI), the clustering was continued to be refined into three subclasses.

Question 4:The same sample in the annex contains features collected at different time points, please analyze them in relation to the time points to uncover patterns in the evolution of different categories of diseases over time.

Question 5:Please consult the relevant literature to describe the early intervention and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI, and AD

2. Problem analysis

2.1 Analysis of question one

To explore the relationship between data characteristics and Alzheimer's disease diagnosis, the data that had been preprocessed were first de-weighted, then processed for missing values, and finally corrected for indicators that deviated from normal values using box line plots. Descriptive statistical analysis was performed on the above preprocessed data to make a stacked histogram to visually analyze the effect of indicators on the diagnosis of Alzheimer's disease and the relationship. The correlation of all indicators with the diagnosis of Alzheimer's disease was analyzed using random forest, and indicator correlation charts were made. 7

2.2 Analysis of question two

The thing to do in this problem is to design an intelligent diagnosis of Alzheimer's disease using additional structural brain features and cognitive-behavioral features. Actually, it is to construct a multiclassification model with inputs based on data of structural brain features and cognitive-behavioral features (x) and outputs of five stages of Alzheimer's disease (y). To build this multi-classification model, the focus is on two aspects, one is the screening of influencing factors and the other is the selection of models. For the screening of features, we can determine the correlation between each influencing factor and Alzheimer's disease based on the weights of the data after processing in the first question, and also combine the existing medical literature and clinical data to determine the final selected influencing factors. Due to the large amount of data and high latitude of this question, it is more appropriate to select a machine learning model. To improve the accuracy of intelligent diagnosis, we select multiple machine learning classifiers to train simultaneously, design each classifier weight according to the classifier performance advantages and disadvantages, and build an integrated learning classifier as the optimal diagnosis model.

2.3 Analysis of question three

Firstly, the baseline dataset was determined as the analysis object, and the important indicators were selected as the clustering features. After verifying the rationality of the classification by elbow diagram, the dataset was then divided into three major categories of CN, MCI and AD by K-means clustering algorithm, and finally the clustering results were visualized and analyzed by PCA downscaling to two-dimensional plane, and the clustering results were obtained by doing another K-means clustering for the MCI category in the above clustering results to divide the dataset into three categories of diseases, SMC, EMCI and LMCI, and visualized and analyzed by PCA downscaling to two-dimensional plane to obtain the subdivided clustering results.

2.4 Analysis of question four

In order to study the evolution law of the same sample at different time nodes, the data samples were first screened, and the data collection scheme was determined for the sake of the time depth of the sample data, searching for indicators closely linked to the time nodes, using the stepwise regression analysis model to screen out the indicators closely linked to time and test the rationality, then normalizing the data indicators by the entropy weight method, drawing a line graph linking time and indicators, and finally doing the time series model to predict the indicator values at the next time node, and analyzing its disease evolution law by the above steps.

2.5 Analysis of question five

Question 5 asked us to review the relevant literature to propose early intervention and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI, and AD, and we wrote a reasonable report combining the papers and considering the above models.

3. Symbol and Assumptions

3.1 Symbol Description

Symbol	Symbol meaning
m_i	Sample Points
x_i	The ith indicator
C_i	The ith cluster
p	The sample point in C_i, m_i is the center of mass of C_i

3.2 Fundamental assumptions

1) The diagnosis of the disease was correct in all participants of the trial.

2) Accurate and reliable sample data with no gross errors.

3) The causes of Alzheimer's disease were all caused by the indicators in the data set, i.e., no other factors were considered to affect Alzheimer's disease.

4. Model

4.1 Problem one modeling and solving

4.1.1 Data preparation and description

The data for this study were obtained from the ANDI database, a global study dedicated to the study and treatment of Alzheimer's disease.

The attachment "ADNIMERGE_New.csv" gives the specific information characteristics (one number at one time point) of 4850 cognitively normal older adults (CN), 1416 patients with subjective memory impairment (SMC), 2968 patients with early mild cognitive impairment (EMCI), 5236 patients with late mild cognitive impairment (LMCI), and 1738 patients with Alzheimer's disease (AD) collected at different time points. specific information characteristics (a time point is a quantity).

The characteristics in this annex can be broadly divided into three main categories: basic information about the participant (age, race, skin color, etc.), baseline physiological data measured when the participant first participated in the experiment and the method used for testing (hereafter referred to as "baseline data"), and physiological data obtained from the test after the participant's condition has evolved over time and the method used for testing (hereafter referred to as "post data").

Since the baseline data and the post data have the same indicators, our team considered that the post data was obtained after a period of treatment or related recovery training, and it does not reflect the most realistic information of the original disease of the participants compared with the baseline data. Therefore, when it comes to the intelligent diagnosis of Alzheimer's disease, we analyze the data set with the basic information of the participants and the baseline data. When considering the pattern of Alzheimer's disease over time, the analysis is combined with the later data to ensure the accuracy and reliability of the analysis.

4.1.2 Random forest feature importance ranking

Random forest [1] is an integrated learning algorithm that uses multiple decision trees for training and makes predictions. To measure the correlation between each data feature in the sample and the diagnosis of Alzheimer's disease, the random forest-based feature importance index (PIM) is calculated for each of the 48 data features according to the equation eq. (1), and the importance of the data features is ranked according to the PIM value, as follows:

①Construct M decision trees;

②When the current decision tree $k_{tree} = 1$, the corresponding out-of-bag data OOB_k is obtained;

③Compute the prediction error $errOOB_k$ of the current decision tree for OOB_k;

④Set the random perturbation of the *i*th data feature in OOB_k to OOB_k^i;

⑤For each decision tree, $k_{tree} = 2,--M$, repeat the steps ②to ④;

⑥Calculate the importance of data features according to equation eq. (1)

$$PIM = \sum_{i}^{M}(errOOB_k^i - errOOB_k) \tag{1}$$

where:M is the number of constructed decision trees, $errOOB_k^i$ and $errOOB_k$ denote the prediction error in the case of k_{tree} decision trees for the out-of-bag data after adding perturbations to the *i*th statistical covariate and the out-of-bag data without adding perturbations, respectively.

4.1.3 Solution of problem one model

4.1.4 Repeat value processing

We were concerned about the occurrence of multiple sample duplicates due to staff errors. Therefore, duplicate value detection was performed first. And the duplicate values were removed.

4.1.5 Missing value handling

We extracted the baseline data from the attachment to obtain the baseline dataset. By analyzing the baseline dataset, we found that several metrics such as AV45, EcogSPTotal, etc. had a missing rate of more than 50%. We queried the official ADNI website for this phenomenon and found that these data with excessive missing rates were new indicators added in the ANDIGO phase, and before that, in the ANDI1 phase, these indicators were not tested, so the data were seriously missing. So we removed the data from ANDI1 stage and finally obtained a data set consisting of 8970 samples from ADNIGO, ADNI2, and ADNI3 stages, which ensured the relative integrity of indicators and data and avoided the impact of too many missing data on the subsequent analysis.

For the processed baseline dataset, we removed metrics such as RID, COLPROT, ORIGPROT, PTID, SITE, EXAMDATE, VISCODE, FSVERSION , FLDSTRENG , MONTH, and M that were not related to Alzheimer's disease. At the same time, to ensure the data interpolation effect, we also excluded the indicators with missing rate greater than 50%, such as PIB, FBB, etc. The final missing rates of each indicator are shown in the following figure.

It can be seen that the missing rate of all indicators is between 0 and 16%, and filling the data above this rate will not destroy the original characteristics of the data excessively.

For missing data of basic physiological indicators, we adopt mean padding, which will not have a large impact on the overall data. As for basic information, such as race, gender, etc., we

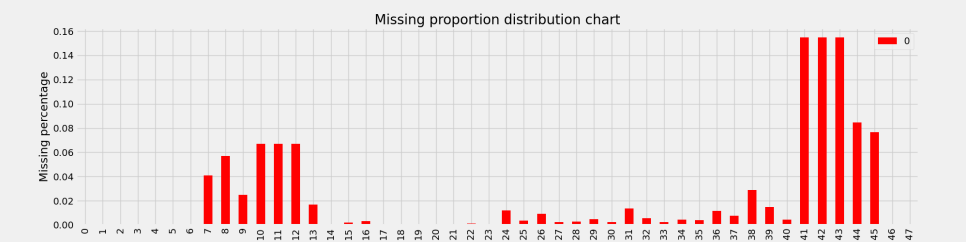


Figure 1 Missing rate control chart

Table 1 Comparison table of data distribution before and after replacement

Before and after replacement	count	mean	std	min	max
Before Replacement	8970	-3.551	5.258	-23.690	7.423
After replacement	8970	-3.101	4.568	-14.960	7.423

take the before-and-after neighboring value filling method, so that the filling strategy appears more random and more realistic.

4.1.6 Outlier handling

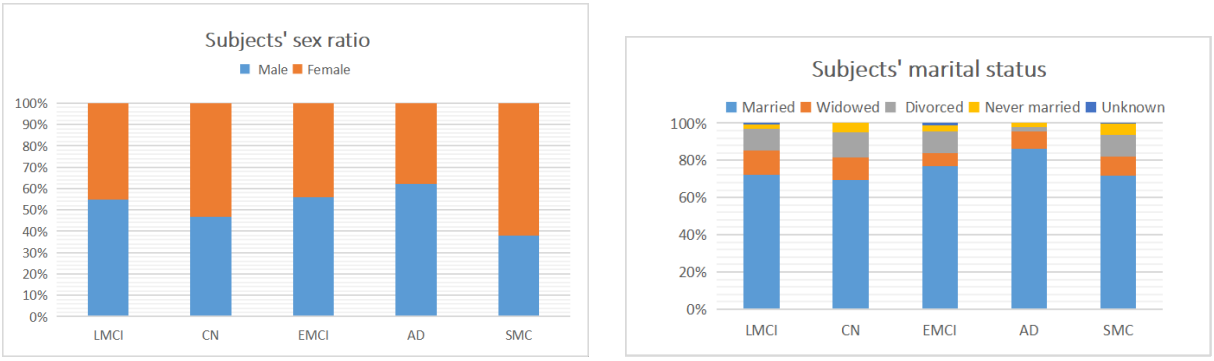
Outliers are those observations that are far from normal, i.e., ”outlier” observations. The presence of outliers can have serious consequences for model creation and prediction. There are two methods to detect outliers, the standard deviation test and the box plot discrimination method. If the data are approximately normally distributed, the standard deviation test is preferred because the data are relatively symmetric, otherwise the box plot discrimination method is preferred. Here, MPACCtrailsB is used as an example for analysis (other indicators are shown in the supporting materials), and both methods are used to detect outliers and plot its corresponding kernel density curve and box plot, as shown below.

As shown in the results of the above run, both the standard deviation discrimination method and the box plot detection method found outliers in the MPACCtrailsB index, and the outliers were all above the lower threshold value. It is obvious from the histogram and kernel density curve that the shape of the data distribution is skewed.

For the detected outliers, the mean value is used for filling, and finally the statistical description before and after the outlier processing is given as shown in Table 1, from which it can be seen that for the replacement of outliers, the mean, minimum, and standard deviation of the original data are changed. And after the change, there is a trend towards the middle, and the overall data distribution is more stable, which is obvious.

4.1.7 Data descriptive statistics

The final dataset we obtained contained 2968 EMCI, 2063 CN, 1610 LMCI, 1416 SMC, and 899 AD with a mean age of 66.9 years. Of these, 4613 were males, accounting for 51.4%, and 4357 were females, accounting for 48.6%. A percentage bar stack plot of gender versus stage of condition was plotted to see the proportion of gender in different condition categories, as shown in Figure. It can be seen that in the CN population, the proportion of males to females is comparable, but in LMCI, EMCI, and AD, the proportion of males is greater than that of females, which may indicate that males are more likely to develop Alzheimer's disease.



(a) Subjects' sex ratio stacked bar graph

(b) Subjects' marital status stacked bar chart

Figure 2 Stacked Bar Chart of Subjects' Sex Ratio and Marital Status

The graph above shows the marital status of people in different condition categories. It can be seen that the marriage rate of CN is at the lowest level, while the marriage rate of SMC, LMCI, EMCI, and AD increases in that order, which may predict that marriage increases the risk of developing Alzheimer's disease.

4.1.8 Random forest analysis correlation

In order to ensure that the importance ranking of indicators has reliability and stability, the maximum decision tree parameter of random forest is set to 5000 through repeated trial calculations, and the PIM values of 47 categories of indicators are calculated according to equation eq. (1), and Figure 3 gives the importance ranking of indicators.

From the figure, it can be concluded that CDRSB, LDEL TOTAL, mPACCdigit, mPACCtrailsB, EcogSPMen and other indicators have a strong correlation with the diagnosis of the disease, all of them are greater than 0.04%. The indicators such as ADAS11, AV45 and others located in the middle of the ranking have some correlation because the correlation degree is mostly the same, located around 0.02%, so they cannot be discarded in the next questions, but remain to be added to the evaluation. As for the last few indicators such as PTRACCAT, PTETHCAT, etc., the correlation is very small and almost tends to 0, so they can be removed.

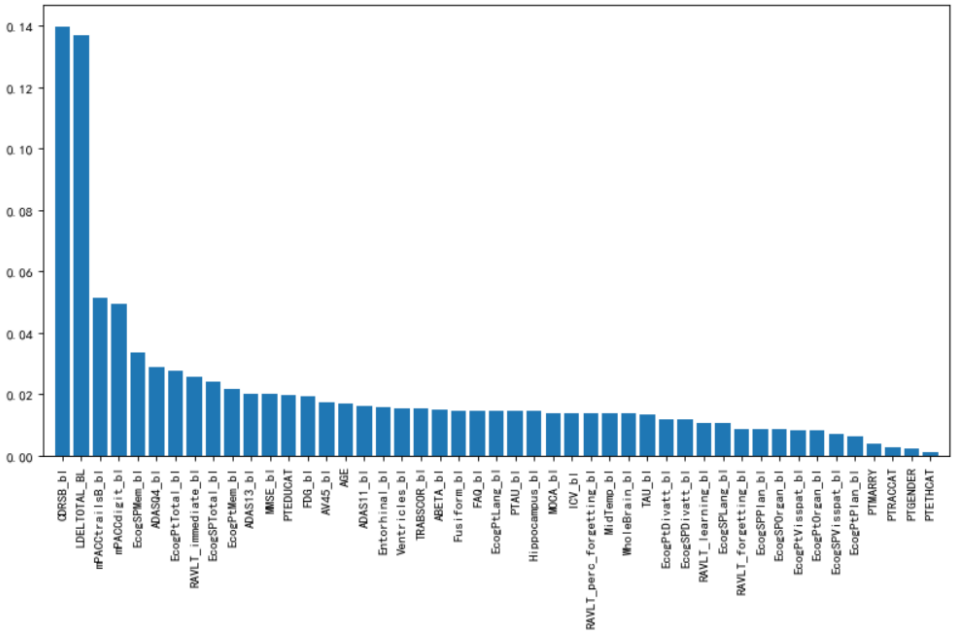


Figure 3 Random Forest Correlation Chart

The four deleted indicators include PTMARRY, PTRACCAT, PTGENDER, and PTETHCAT.

4.2 Problem two modeling and solving

The problem requires the design of intelligent diagnosis [2] for Alzheimer's disease based on the structural brain features and behavioral structural features in the ADNI library data. This belongs to the five classification problem solving, which can be divided into two sub-problems, one is feature selection and processing, and the other is constructing the intelligent diagnosis classification model. Firstly, the data indicators are screened based on the weights of each feature indicator derived from the random forest algorithm based on the preprocessing results in the first step, and the definition of structural brain features and cognitive behavioral features by medical expert knowledge. Then the data were divided into a training set and a validation set (test set). Next, three models were built, logistic regression model, xgboost model, and llightgbm model. A combination of smote+enn sampling was used to handle the uneven data and cross-validation was added. Based on the classification results, prediction results, confusion matrix, and roc curves, the performance strengths and weaknesses of the classifiers are judged and improvements are made. Finally, we design the weights of each classifier, build an integrated learning classifier as the optimal diagnosis model, and establish an intelligent diagnosis system of Alzheimer's disease based on multi-model fusion. The process of establishing the intelligent diagnosis system is shown in the following figure.

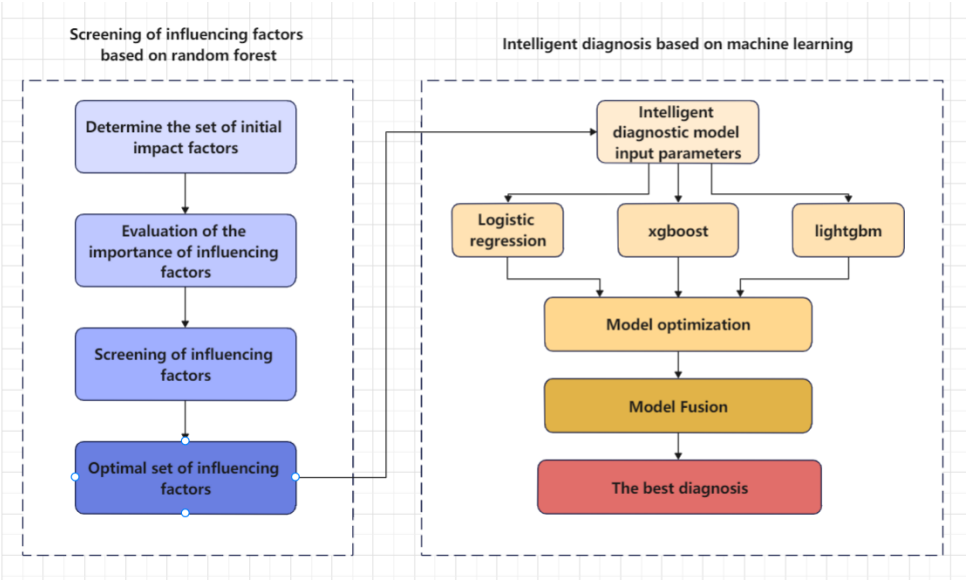


Figure 4 An intelligent diagnosis system for Alzheimer's disease based on multi-model fusion

4.2.1 Logistic regression model

Logistic regression [3](Logitc Regression) is a regression model proposed by statistician David Cox in 1958, in fact, logistic regression algorithm is a classification algorithm that can be used to deal with binary and multi-classification problems, and more commonly used in binary classification problems, it is mainly Classification of the sample space through the logistic function. The currently commonly used logistic function sigmoid function (for the convenience of representation, it will be noted as σ, defined as follows.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2}$$

The output values of linear regression are a large range of numbers, e.g., from negative infinity to positive infinity, while logistic regression compresses the values of linear regression to between (0,1) by Suppose the data set for the binary classification is $\{a_i, y_i\}_{i=1}^{m}$, a_i is the n-maintained positive vector, $y_i \in \{0,1\}$ is the label of the sample, and let x be the parameters of the model, i.e., the regression coefficients, then the probabilities of y=1 and y=0 can be expressed as

$$P(y=1 \mid a,x) = p(a) = \frac{1}{1+e^{-(x)|a}} \tag{3}$$

$$P(y=0 \mid a,x) = 1 - p(a) = 1 - \frac{1}{1+e^{-(x)^{\top}a}} = \frac{e^{-(x)^{\top}a}}{1+e^{-(x)^{\top}a}} = \frac{1}{e^{-(x)^{\top}a}} \tag{4}$$

This is equivalent to $\boldsymbol{P}(y \mid a,x) = p(a)^{y}(1-p(a))^{1-y}$. The most basic learning algorithm of logistic regression, maximum likelihood estimation, is based on the basic principle of observing the outcome of a known sample through several experiments, thus inferring the parameter values

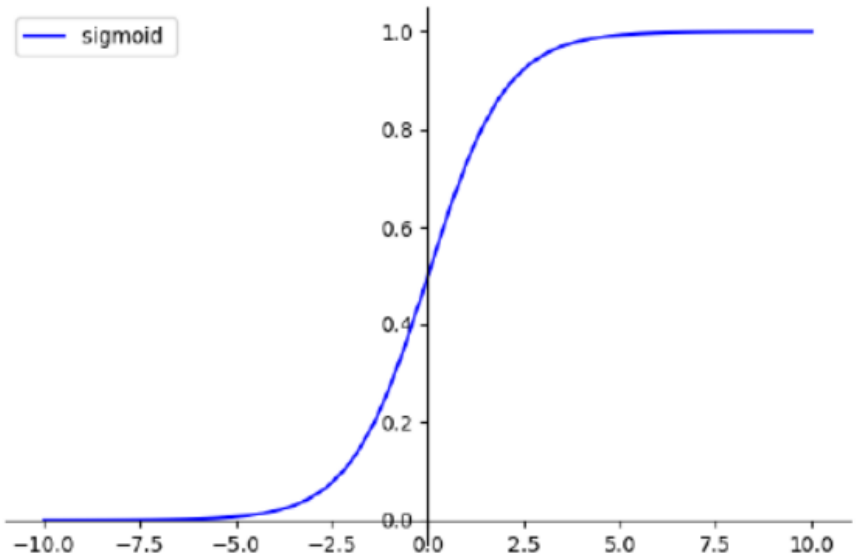


Figure 5 sigmoid function

that lead to such an outcome with maximum probability, and when the samples are independent, the likelihood function is

$$I(x) = \prod_{i=1}^{m} P\left(y_i = 1 \mid a_i, x\right)^{y_i} \left[1 - P\left(y_i = 1 \mid a_i, x\right)\right]^{1-y_i} \tag{5}$$

Taking logarithms on both sides of the above equation, the log-likelihood function can be obtained as

$$\begin{aligned} L(x) &= \ln(l(x)) \\ &= \sum_{i=1}^{m} y_i \ln\left(p\left(a_i\right)\right) + \left(1 - y_i\right) \ln\left(1 - p\left(a_i\right)\right) \\ &= \sum_{i=1}^{m} y_i \left[\ln\left(p\left(a_i\right)\right) - \ln\left(1 - p\left(a_i\right)\right)\right] + \ln\left(1 - p\left(a_i\right)\right) \\ &= \sum_{i=1}^{m} y_i x^T a_i - \sum_{i=1}^{m} \ln\left(1 + e^{x^T a_i}\right) \end{aligned} \tag{6}$$

The derivative of $L(x)$ with respect to x yields:

$$\begin{aligned}\frac{\partial L(x)}{\partial x} &= \sum_{i=1}^{m} y_i a_i - \sum_{i=1}^{m} \frac{e^{x^T a_i}}{1 + x^T a_j} a_j \\ &= \sum_{i=1}^{m} y_i a_i - \sum_{i=1}^{m} \frac{1}{1 + x^T a_j} a_j \\ &= \sum_{j=1}^{m} \left(y_i - \frac{1}{1 + x^T a_j} \right) a_i \\ &= \sum_{i=1}^{m} \left[y_i - \sigma \left(x^T a_i \right) \right] a_j \end{aligned} \tag{7}$$

4.2.2 lightbgbm model

The lightbgbm algorithm is an improved version of the GBDT algorithm proposed by the Microsoft team in 2017. compared to XGBoost, which is also based on the improved GBDT algorithm, LightGBM solves some of the shortcomings of XGBoost.

(1) lightbgbm uses the Histogram algorithm. the XGBoost algorithm uses a pre-sorting algorithm, which first sorts the samples according to the feature values and then finds the optimal splitting points from all the feature values. the number of candidate splitting nodes in this algorithm is proportional to the number of samples. The Histogram algorithm used by lightbgbm reduces the number of candidate split nodes by discretizing the continuous eigenvalues into a fixed number of bins, so that the number of candidate split nodes is constant.

(2) lightbgbm uses the GOSS algorithm: Gradient-based One-Side Sampling, which is known as Gradient-based One-Side Sampling. The main idea is to reduce the complexity of calculating the gain of the objective function by sampling the samples, and this algorithm strikes a balance between computational performance and computational accuracy.

(3) lightbgbm uses the EFB algorithm: a mutually exclusive feature bundling algorithm. the EFB algorithm solves the problem that XGBoost generates a large number of sparse features due to the use of one-hot encoding.

4.2.3 xgboost model

The XGBoost [4] integration algorithm mainly uses a tree structure and constantly performs feature splitting to grow a new tree. That is, XGBoost continuously generates new decision trees A, B, C, D... The final algorithm of the generated decision tree is the sum of $A + B + C + D + ...$ of the sum of the decision trees. One function is added at a time to fit the error predicted by the previous layer. To prevent overfitting, XGBoost introduces an $L2$ regular term to smooth the predicted values of the leaf nodes. The objective function consists of a loss function and a regularization term. Its objective function is :

$$L(\varphi) = \sum_i l\left(y_i, y_i\right) + \sum_k \Omega\left(f_k\right) \tag{8}$$

The regularization term is:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{9}$$

Where T represents the number of leaf nodes and ω denotes the fraction of leaf nodes. The regularization term represents a function of the complexity of the tree, and the smaller the value, the lower the complexity and the better the generalization ability. xGBoost allows the use of cross-validation in each boosting iteration round. Therefore, the optimal number of boosting iterations can be easily obtained. xGBoost uses a greedy algorithm to enumerate all possible segmentation processes on all features, and its flow is shown below:

Algorithm 1 Exact Greedy Algorithm for Split Finding

Input:1,instance set of current node
Input:d,feature dimension
$gain \leftarrow 0$
$G \leftarrow \sum_{i\in I} g_i, H \leftarrow \sum_{i\in I} h_i$
for k=1 to m do
$G_L \leftarrow 0, H_L \leftarrow 0$
    for j in sorted (I,by,X_{ik}) do
        $G_L \leftarrow G_L + g_{ui}, H_L \leftarrow H_L + h_{vi}$
        $G_R \leftarrow G - G_L, H_N \leftarrow H - H_L$
        score←-Score $= \max\left(\text{score}, \frac{1G_L^2}{2H_L \mid \lambda} + \frac{1G_R^2}{2} - \frac{1}{H_R \mid \lambda} - \frac{(G_L \mid G_R)^2}{2} - \gamma\right)$
    end
end
Output:Split with max score

Figure 6 Exact Greedy Algorithm for Split Finding

The partial tree diagram after xgboost visualization is shown in Fig:

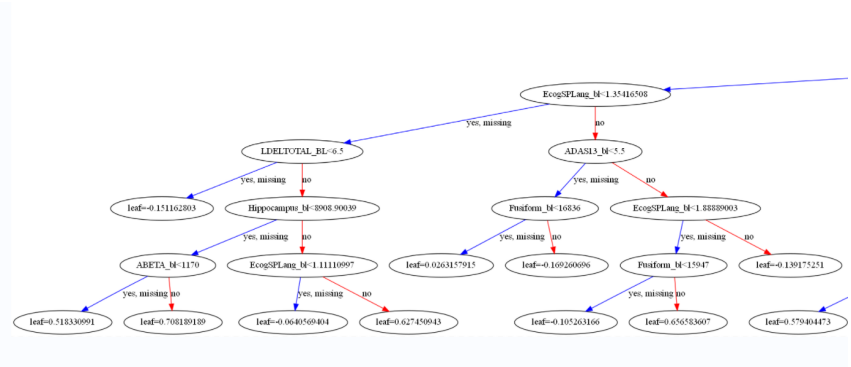


Figure 7 xgboost partial tree diagram

4.2.4 Solution of problem

4.2.5 Data Metrics Filter

By examining the provided ADNI data, we found that it contains two time samples, and the diagnosis of Alzheimer's disease in the second time sample is only divided into three stages, CN, MCI, and AD. and many of them are obviously not related to brain characteristics and cognitive behaviors, such as (race, marriage, etc.). We considered that if we use the original dataset directly for training, the accuracy of the diagnostic model obtained may not be very high, and it will also increase the time required for model training. Therefore, we decided to first filter the indicators by the random forest algorithm based on the pre-processed data in the first question, and also combined with the medical experts' knowledge on the definition of structural brain features and cognitive-behavioral features, and finally obtained the bar chart of the filtered indicator feature weights, as follows.

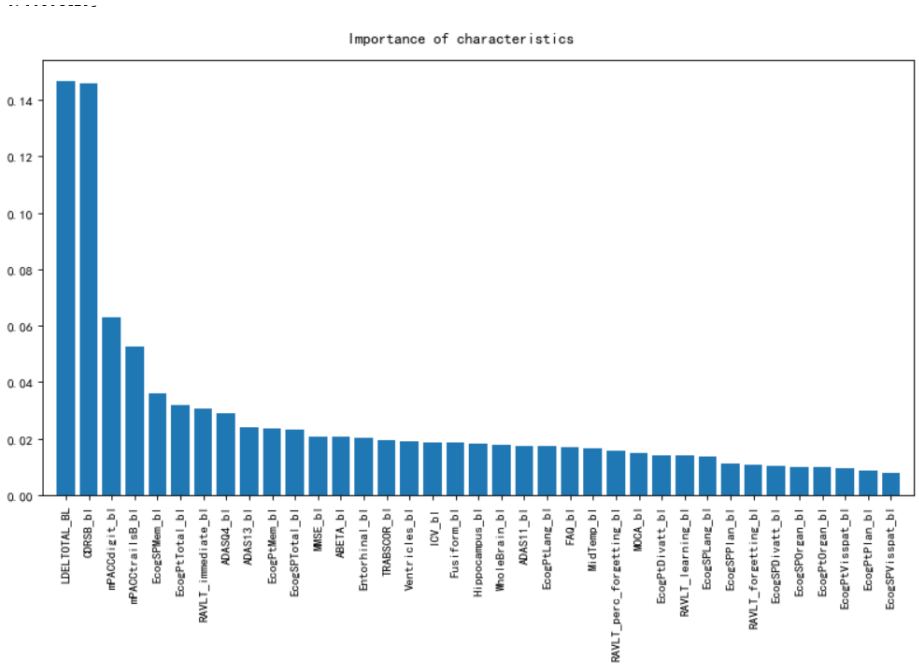


Figure 8 Bar Chart of Indicator Feature Weights

4.2.6 Training and Prediction

After getting the screening completed influencing factors, we use logistic regression model in sklearn that provides classification and regression, xgboost model, lightgbm model, call Classifier function, input to brain structural features and cognitive behavioral features data (x), output Alzheimer's disease five stages (y) for machine learning. And the training set and test set were divided according to a certain ratio, and the data were processed again by combined sampling smooe+enn, and the model was evaluated on the training set using 5-fold cross-validation, and

the fi-core average was taken as the final score of the model. The specific results of the classification are shown in Table.

Classifier	Test set accuracy	Accuracy of cross-validation set
Logistic	33%	29.40%
lightgbm	99%	98.77%
Xgboost	98%	98.63%

Figure 9 Classification results

We found that the classification models based on the xgboost algorithm and lightgbm were highly accurate in diagnosing the five stages of Alzheimer's disease development, both in terms of accuracy on the test set and in terms of the mean of the 5-fold cross-validation, and the difference in the cross-validation results reflected the superior fitting ability of the integrated machine learning method over the traditional machine learning models.

4.2.7 Results and Analysis

In order to analyze the advantages and disadvantages of the three types of machine learning models more intuitively and thus determine the diagnostic model, we analyzed the confusion matrix, ROC plot, by plotting the xgboost, lightgbm, and logistic regression models in combination with the prediction results.

In classification problems, the results of the classification are usually visualized and analyzed using a confusion matrix, which gives a very clear picture of how well the classification model identifies the various Alzheimer's developmental stages. The figure below shows the confusion matrix for the classifier we used on the test set.

Through the confusion matrix of the three models shown below, it can be seen that on the dataset labeled with CN, SMC, EMCI, LMCI, and AD, the two models xgboost and lightgbm have the best training effect, with the accuracy of classification of all kinds of indicators reaching more than 95%, and the proportion of misclassification is minimal. In contrast, the performance of logistic regression models on this dataset was not satisfactory, and the accuracy of PAtt3 (EMCI), which had the best classification effect, was only just over 50%.

ROC curve refers to receiver operating characteristic curve (receiver operating characteristic curve), which is a comprehensive index reflecting the sensitivity and specificity of continuous variables, and the interrelationship between sensitivity and specificity is revealed by graphical method. The larger the area under the curve, the higher the diagnostic accuracy. The larger the

(a) Logistic regression

(b) xgboost

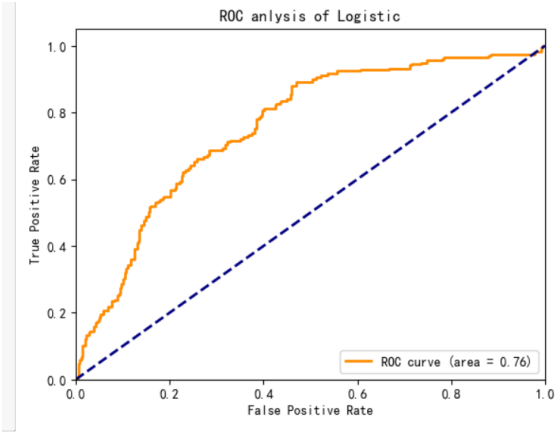
(c) lightgbm

Figure 10 Confusion Matrix

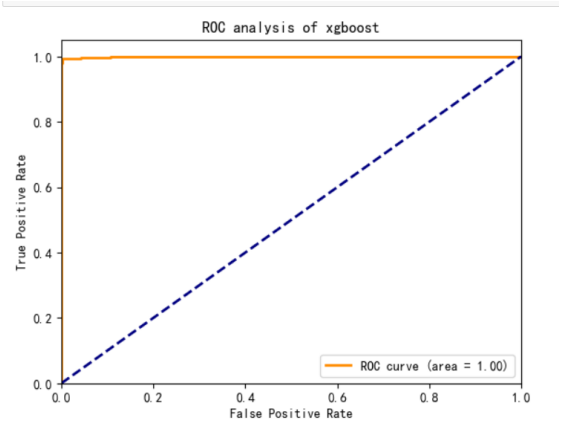
area under the curve, the higher the diagnostic accuracy. The point closest to the top left of the ROC curve is the threshold value with higher sensitivity and specificity. The ROC curves for the three models are shown below:

From the above graph we can see that the classification accuracy of xgboost and lightgbm is approaching 1, which is much higher than the logistic regression classification accuracy. Comparing the roc curves of xgboost and lightgbm, lightgbm grows faster than xgboost in the vertical coordinate and reaches 1 earlier, but by and large there is almost no difference in the area below the curve. Although the results are very good, the possibility of overfitting cannot be ruled out, because based on the characteristics of the lightgbm and xgboost models, overfitting often occurs easily when the sample size is not uniform and the noise is too large.

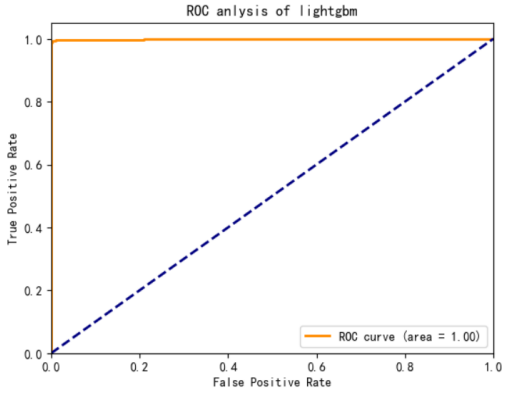
After analyzing the prediction results of the three types of machine learning, we found that



(a) Logistic regression



(b) xgboost



(c) lightgbm

Figure 11 ROC graph

the logistic regression model has a higher diagnostic accuracy for the SMC stage up to 60%, but a lower diagnostic accuracy for other stages especially for the first stage, while for the xgboost model, the diagnostic accuracy for the first stage CN is as high as 98% other stages are weaker, and the lightgbm model has a better diagnostic accuracy for the first stage CN The accuracy of the lightgbm model for the diagnosis of the first stage CN is not as good as the xgboost model, but the accuracy of the diagnosis in the AD stage performs better. The prediction results are as follows:

Considering the differences in the diagnostic accuracy of the above three models at each stage of Alzheimer's disease, we decided to use a voting integration strategy to combine three classifiers, logistic regression, xgboost, and lightgbm, according to certain weights, to obtain an integrated learning classifier. To compensate for the deficiency of a certain machine learning model in diagnostic accuracy for a certain stage of Alzheimer's disease. The final 98% was obtained on the test set and 97.89% after cross-validation.

In summary, the intelligent diagnosis system of Alzheimer's disease based on multi-model

CN	SMC	EMCI	LMCI	AD
0.01226589	0.675320949	0.213058994	0.097120226	0.00223394
0.029661405	0.637967093	0.195508189	0.13049552	0.006367793
0.029661405	0.637967093	0.195508189	0.13049552	0.006367793
0.04484137	0.559815657	0.229377827	0.152442998	0.013522149
0.04484137	0.559815657	0.229377827	0.152442998	0.013522149
0.021826005	0.534298417	0.29522897	0.141235915	0.007410693
0.021826005	0.534298417	0.29522897	0.141235915	0.007410693
0.021826005	0.534298417	0.29522897	0.141235915	0.007410693
0.038403715	0.502612789	0.243731804	0.201277974	0.013973719
0.038403715	0.502612789	0.243731804	0.201277974	0.013973719
0.121131097	0.497105259	0.132887835	0.220826333	0.028049477
0.056596239	0.487665526	0.245230068	0.187046049	0.023462117
0.056596239	0.487665526	0.245230068	0.187046049	0.023462117
0.101917833	0.479806927	0.217701265	0.178646825	0.021927149
0.101917833	0.479806927	0.217701265	0.178646825	0.021927149
0.102252308	0.476861638	0.171324896	0.225523746	0.024037411

(a) Logistic regression

CN	SMC	LMCI	LMCI	AD
0.99924684	0.000102736	0.000114886	0.000133377	0.000402192
0.99924684	0.000102736	0.000114886	0.000133377	0.000402192
0.99924684	0.000102736	0.000114886	0.000133377	0.000402192
0.99924684	0.000102736	0.000114886	0.000133377	0.000402192
0.99924684	0.000102736	0.000114886	0.000133377	0.000402192
0.99924684	0.000102736	0.000114886	0.000133377	0.000402192
0.99914896	0.000102991	0.000111846	0.000133708	0.000502507
0.99914896	0.000102991	0.000111846	0.000133708	0.000502507
0.999079	0.000120359	0.0001043	0.000124688	0.000571696
0.999079	0.000120359	0.0001043	0.000124688	0.000571696
0.999079	0.000120359	0.0001043	0.000124688	0.000571696
0.999079	0.000120359	0.0001043	0.000124688	0.000571696
0.9990734	0.000109111	0.000102469	0.000119651	0.000595352
0.9990734	0.000109111	0.000102469	0.000119651	0.000595352

(b) xgboost

CN	SMC	EMCI	LMCI	AD
0.868791014	0.000311758	0.000570858	0.0086745	0.121651869
0.868791014	0.000311758	0.000570858	0.0086745	0.121651869
0.868791014	0.000311758	0.000570858	0.0086745	0.121651869
0.866160626	0.00050349	0.000570858	0.0086745	0.124090525
0.866160626	0.00050349	0.000570858	0.0086745	0.124090525
0.866148714	0.000505854	0.000570858	0.00976071	0.123013864
0.866148714	0.000505854	0.000570858	0.00976071	0.123013864
0.865611156	0.000301342	0.000570858	0.0086745	0.124842144
0.865611156	0.000301342	0.000570858	0.0086745	0.124842144
0.865611156	0.000301342	0.000570858	0.0086745	0.124842144
0.865611156	0.000301342	0.000570858	0.0086745	0.124842144
0.865611156	0.000301342	0.000570858	0.0086745	0.124842144

(c) lightgbm

Figure 12 Predicted results

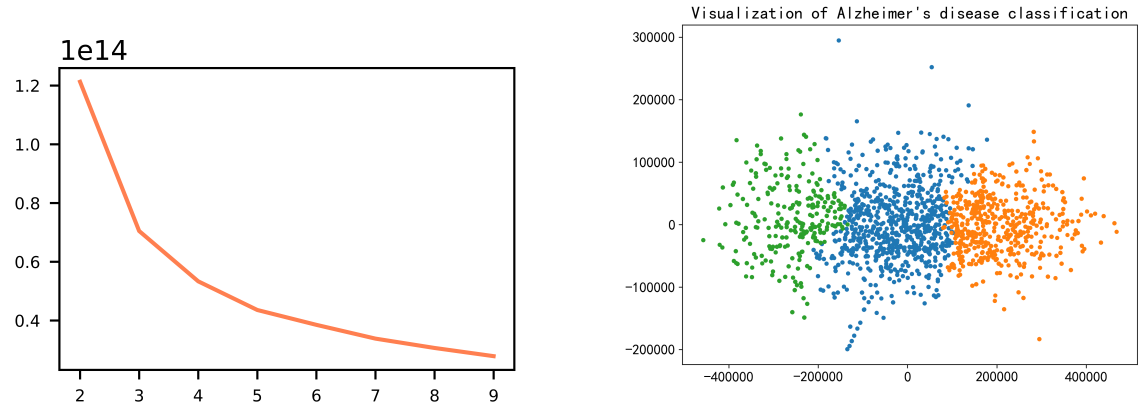
fusion has achieved the expected level for the diagnosis of CN, SMC, EMCI, LMCI and AD. The advantages of multi-models are fully exploited to subtly make up for the shortcomings of a single model and avoid the problem of absolute evaluation of a single model, and the idea is worthy of application and promotion.

4.3 Problem three modeling and solving

4.3.1 Classification of CN, MCI, AD

The baseline data set is taken as the object of analysis, and the indicators not deleted in the first question are selected as the clustering features. Firstly, we verify the rationality of the selected indicators, and draw an elbow diagram as shown in Figure . Through the elbow diagram, we can see that when the number of clusters exceeds 3, the decreasing trend of the graphical fold slows down significantly, so it is most appropriate to classify Alzheimer's disease into three categories, which coincides with the requirement of the question to classify Alzheimer's disease into CN, MCI, and AD, and can prove from the side that we This is in line with the requirement of the question to classify Alzheimer's disease into CN, MCI and AD.

K-means clustering was performed on the data to classify Alzheimer's disease into CN, MCI,



(a) CN, MCI, AD elbow chart

(b) CN, MCI, AD classification chart

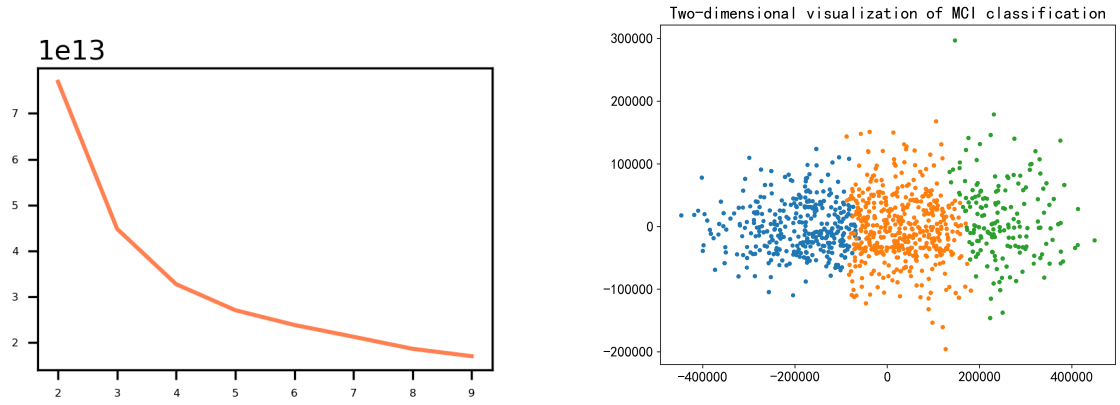
Figure 13 Classification of CN, MCI, AD display chart

and AD categories according to their conditions. Since the data in this question are of high dimensionality and cannot be visualized directly, PCA was used to reduce the dimensionality to a two-dimensional plane in order to perform visualization and analysis. The results showed that the classification effect was reasonable and clear, and the points clustered into the same category had smaller internal distance and were more closely clustered; however, at the boundaries of different types of classification points, CN, MCI, and AD all overlapped to different degrees, which reflected the obscurity of the development of Alzheimer's disease and the difficulty of discriminating each disease stage.

4.3.2 Refine the classification of MCI

The process of refining the classification of MCI was broadly similar to the classification of Alzheimer's disease into CN, MCI, and AD, and we continued to select the indicators not deleted in the first question as clustering features. However, the samples with the disease categorized as SMC, EMCI, and LMCI in the baseline data set were extracted for analysis. The elbow diagram is drawn as shown in Figure. The elbow diagram shows that when the number of clusters exceeds 3, the decreasing trend of the graphical fold slows down significantly, so it is most appropriate to classify MCI into three categories, which coincides with the question's requirement to classify MCI into SMC, EMCI, and LMCI.

The K-means clustering results were also downscaled to a two-dimensional plane for visualization and analysis using PCA downscaling. As shown above, the results of classifying Alzheimer's disease into SMC, EMCI, and LMCI are more discrete within the data than those of classifying Alzheimer's disease into CN, MCI, and AD, and the results of classifying MCI into SMC, EMCI, and LMCI are more overlapping at the classification boundaries, which is consistent with the results we found on the official website of ADNI: i.e., the progression of



(a) MCI Refined Elbow Chart

(b) MCI Refined Classification Chart

Figure 14 MCI refinement classification display chart

MCI is slow and long, and the SMC, EMCI, and LMCI under MCI also partially overlap in physiological data. This is consistent with the results we found on the ADNI website: that is, the progression of MCI is slow and long, and the SMC, EMCI, and LMCI categories under MCI also partially overlap in terms of physiological data, so MCI cannot be classified only from a single dimension. In the actual diagnosis, detailed tests are often required to ensure the accuracy of the classification results.

4.4 Problem four modeling and solving

The questions require the inclusion of features at different points in time for the same sample and the analysis of the relationship between the points in time and these features, and the description of the evolutionary patterns of different types of diseases according to time.

In order to find the indicators that are more related to time at different time points, the data were preprocessed, and in order to facilitate the observation of the evolutionary pattern of different diseases, the final data collection scheme was selected as ADNI3, and the initial collection scheme was ADNI2 as data sample points with a total of 166 sample points, and then screened out indicators such as gender, race, and baseline correlation (other than baseline diagnosis), and in order to better identify the indicators that are related to In order to better identify time-related indicators, the association between screened indicators and time was explored by establishing a stepwise regression analysis model with the discriminant coefficient R^2 as the test quantity, and the degree of integrated association of screened indicators to a single time indicator.

4.4.1 Stepwise regression analysis model

The basic idea of the stepwise regression algorithm is to introduce all the factors into the regression equation one by one according to their influence on the explanatory variables, from the largest to the smallest, and to test all the variables contained in the regression equation at any time to see if they are still significant, and to eliminate them if they are not significant, until all the variables contained in the regression equation have a significant effect on the explanatory variables before considering the introduction of new variables. Then, among the remaining unselected factors, the one with the greatest effect on the explanatory variable is selected and tested for significance, and if it is significant, it is introduced into the equation, and if not, it is not introduced. Until finally there are no more significant factors to introduce and no more insignificant variables to eliminate.

For a determined data series time y and $x_n(x = 1, 2, ..., n)$, the running process of the stepwise regression algorithm can be described by the following steps are described.

Step1 Calculate the index mean $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n, \bar{y}$ and the sum of difference squares L_{11}, L_{22}, $\cdots, L_{pp}, L_{yy}$.

Their respective normalized variables are

$$L_{11}, L_{22}, \cdots, L_{pp}, L_{yy} u_j - \frac{x_j - \bar{x}_j}{\sqrt{L_{ij}}}, j - 1, \ldots, p, \quad u_{p+1} - \frac{y - \bar{y}}{\sqrt{L_{yy}}} \tag{10}$$

Step2 Calculate the correlation coefficient matrix $R^{(0)}$ of $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p$ and y.

Step3 Suppose k indicators have been selected, $x_{i_1}, x_{i_2}, \cdots, x_{i_k}$, and $i_1, i_2, \cdots, i_k$ are different from each other.

$R^{(0)}$ is obtained after transforming

$$R^{(k)} = \left(r_{i,j}^{(k)}\right) \tag{11}$$

For $j = 1, 2, \cdots, k$ calculate the partial regression sum of squares of the standardized variables u_{i_j} one by one

$$V_{i_j}^{(k)} = \frac{\left(r_{i_j,(p+1)}^{(k)}\right)^2}{r_{i,j,j}^{(k)}} \tag{12}$$

Denote $V_l^{(k)} = \max\left\{V_{i_j}^{(k)}\right\}$, as F test

$$F = \frac{V_l^{(k)}}{r_{(p+1)(p+1)}^{(k)}/(n - k - 1)} \tag{13}$$

For a given significance level α, the rejection domain is

$$F < F_{1-\alpha}(1, n-k-1) \tag{14}$$

Step4 Loop through Step3 until finally t variables $x_{i_1}, x_{i_2}, \cdots, x_{i_1}$, and $i_1, i_2, \cdots, i_t$ are not the same as each other.

$R^{(0)}$ after transformation gives

$$R^{(k)} = \left(r_i^{(k)}\right) \tag{15}$$

Then the corresponding regression equation is obtained

$$\frac{\hat{y}-\bar{y}}{\sqrt{L_{yy}}} = r_{i_1,(p+1)}^{(k)} \frac{x_{i_1}-\bar{x}_{i_i}}{\sqrt{L_{i_{i,i}}}} + \cdots + r_{i_k,(p+1)}^{(k)} \frac{x_{i_k}-\bar{x}_{i_k}}{\sqrt{L_{i_k,i_k}}} \tag{16}$$

The final result by algebraic operation is

$$\hat{y} = b_0 + b_{i_1} x_{i_1} + \cdots + b_{i_k} x_{i_k} \tag{17}$$

4.4.2 Evolutionary law analysis model based on entropy weight method and time series

Through the above model, the time-related indicators can be filtered out from many indicators, and then the indicators can be downscaled by the entropy weight method[7], which is established in the following steps.

Step1 Construct the decision matrix of the components.

$$X_{m\times n} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ & & \vdots & \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{18}$$

Step2 Regularization process

$$\lambda_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \tag{19}$$

Step3 Calculate the weight of the ith value under the jth item

$$p_{ij} = \frac{\lambda_{ij}}{\sum_{i=1}^{m} \lambda_{ij}} \tag{20}$$

Step4 Calculate the entropy value of each component

$$e_j = -k \sum_{i=1}^{m} p_{ij} \ln\left(p_{ij}\right) \tag{21}$$

where K is $\frac{1}{In(m)}$

Step5 Calculate the weighting factor

$$\omega_j = \frac{d_j}{\sum_{i=1}^{n} d_j} \times 100\% \tag{22}$$

Based on the above conclusion, a new indicator can be defined as y_i, which is calculated as

$$y_i = \sum_{i=0}^{m} \omega_i x_i \tag{23}$$

The above new indicator y_i is substituted into the following autoregressive ARMA model [8] If the time series y satisfies.

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + \varepsilon_t \tag{24}$$

x_t is the value of the time series x at moment t, $\phi_1, \phi_2 \ldots, \phi_p$ are the autoregressive coefficients, and ε_t is a series of independent identically distributed random variables satisfying.

$$E\left(\varepsilon_t\right) = 0, \mathrm{Var}\left(\varepsilon_t\right) > 0 \tag{25}$$

Then the time series x is said to obey a p-order autoregressive model $AR(p)$.

4.4.3 Solution of problem four model

The pre-processed data were first visualized by combining the data from the benchmark diagnoses (AD, CN, EMCI, LMCI, SMC) with the diagnostic data (CN, Dementia, MCI), respectively, using SPSS to the bar chart as shown in fig. 15 below.

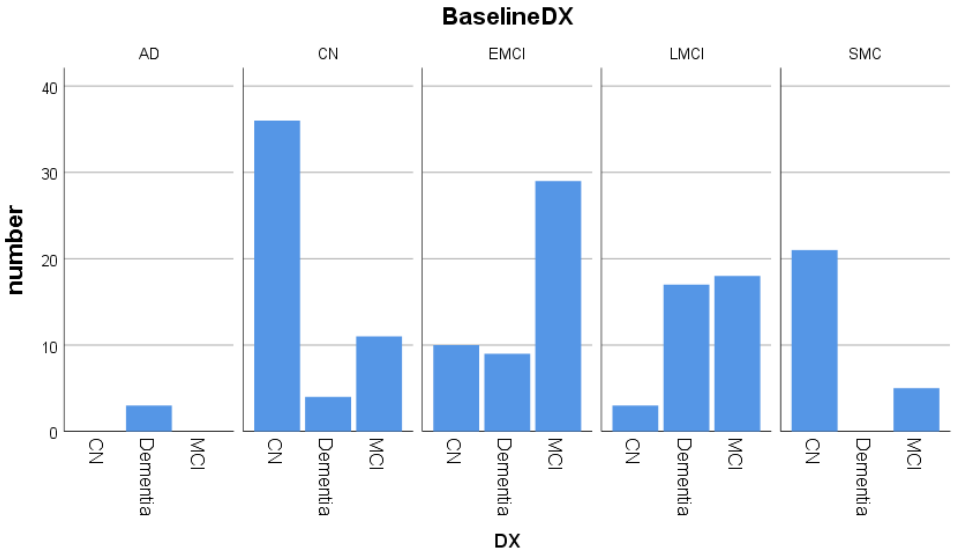


Figure 15 Bar chart of BaselineDx vs Dx

As can be seen from the above graph, there is a certain time evolution of the disease. For people with a baseline diagnosis of CN, most of them will have CN after a few years, and some of them will have MCI or Dementia. For those with a baseline diagnosis of SMC, most of them will return to normal after a few years, while a small number of them will deteriorate to MCI.

4.4.4 Stepwise regression analysis model solving

With time as y and other indicators as x_i, the stepwise regression model is solved using SPSSPro, and after n=57, the calculated results are shown in table 2.

Table 2 Stepwise regression analysis of discarded and retained indicators

Indicators	Results of stepwise regression analysis
Abandonment of indicators	DIGITSCOR,EcogPtLang,TRABSCOR,mPACCdigit, EcogPtMem,Fusiform,MidTemp,ICV,EcogPtVisspat, EcogPtDivatt,EcogSPMem,EcogSPLang,EcogSPVisspat, EcogSPOrgan,EcogSPDivatt,EcogSPTotal,IMAGEUID WholeBrain,Entorhinal,MOCA,FAQ,RAVLT_perc_forgetting, CDRSB,RAVLT_learning,RAVLT_immediate,MMSE,, ADASQ4,ADAS13,FBB,ABETA,ADAS11,PTAU,TAU,AV45,PIB mPACCtrailsB,FDG,EcogPtPlan,EcogPtOrgan,EcogSPPlan ,Ventricles,Hippocampus,RAVLT_forgetting
Retention Indicators	EcogPtTotal,EcogSPTotal

The above results show that the indicators EcogPtTotal and EcogSPTotal have a greater association with time.This paper also conducts F-test on the indicators through SPSSPro and calculates the discriminant coefficient R^2, and the test results are as follows

Table 3 Stepwise regression analysis test table

Indicators	t	p	R^2	AdjustmentR^2	F
EcogPtTotal	-4.594	0.000***	0.669	0.657	0.000***
EcogSPTotal	3.942	0.000***	0.564	0.556	0.000***

Note: ***, **, * represent the significance level of 1%, 5%, 10%, respectively

According to the results obtained above, the relationship between time t and indicators EcogPtTotal and EcogSPTotal is very strong, and the effect of time t on indicators EcogPtTotal and EcogSPTotal is correlated significantly at the 0.01 level, then the indicators EcogPtTotal and EcogSPTotal are extracted as the observation and analysis of the same The main indicators for observing and analyzing the disease evolution pattern of the same sample at different times were extracted.

4.4.5 Evolutionary law analysis based on entropy method and time series

The indicators EcogPtTotal and EcogSPTotal were extracted from the above findings, and in order to better see the effect of time variation on the indicators, the indicators were normalized, and using SPSSPro, the indicators EcogPtTotal and EcogSPTotal were processed using the entropy weighting method, and the results of the weighting analysis are shown in the following table table 5.

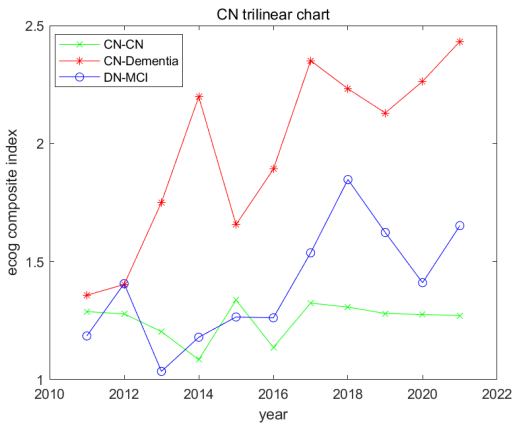
Table 4 Calculated results of weight analysis based on entropy weighting method

Indicators	Information entropy value e	Information utility value d	Weights
EcogPtTotal	0.963	0.037	36.675%
EcogSPTotal	0.937	0.063	63.325%

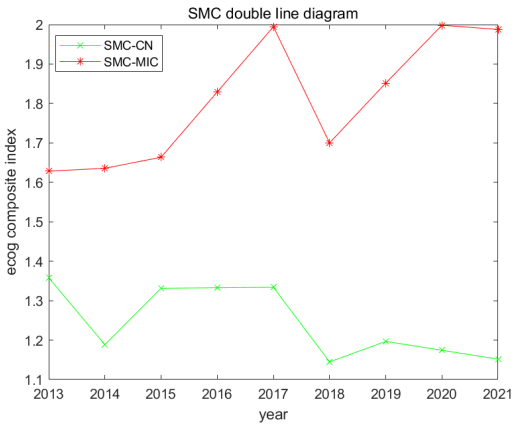
Through the above entropy weighting method weighting analysis calculation results, we can get the Ecog sum judging index $y = 0.36675 * EcogPtTotal + 0.63325 * EcogSPTotal$, in order to better observe the change pattern of the disease, based on the results of the benchmark diagnosis and ANDI3 final diagnosis results for classification, for example, the benchmark diagnosis results for CN The samples with the final diagnosis of ANDI3 as CN result were grouped into one category, a total of 12 categories, and then 3 samples were randomly selected from multiple sample points of each category, and the mean value of the drawn sample points y versus time was made into a line graph using matlab as follows.

The above line graph illustrates that for the CN group, there was no significant difference at the beginning, and over time, there was a continuous fluctuating increase for those who were eventually diagnosed with Dementia and MCI, with a greater increase in indicators for those who were eventually diagnosed with Dementia. For the SMC group, there was a significant difference between the final diagnosis of CN and MCI. For the LMCI and EMC groups, those with a final diagnosis of CN showed a fluctuating decrease in indicators, those with a final diagnosis of Dementia showed a fluctuating increase in indicators, those with a final diagnosis of MCI were more stable, and for the AD group, there was no significant change in indicators.

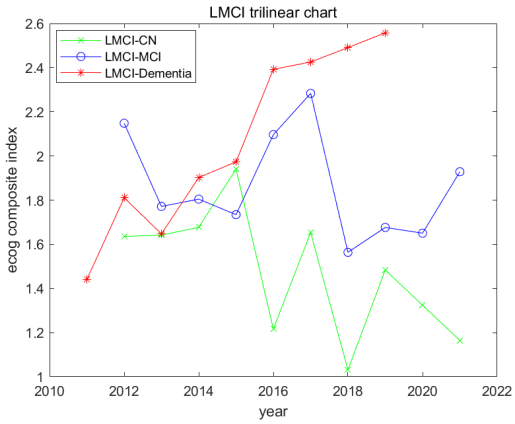
Finally, in order to visualize the pattern of disease changes, a time series extrapolation was



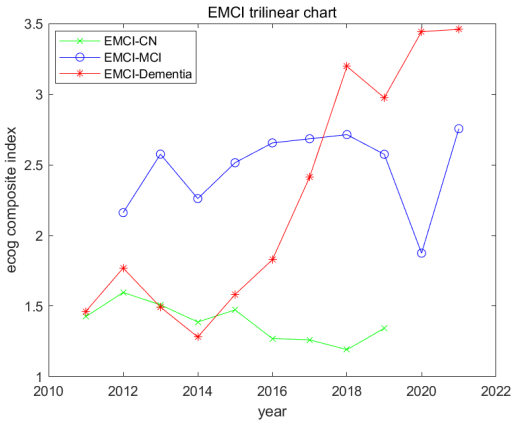
(a) CN vs CN, MCI, Dementia fold trend graph



(b) Folding Trend of SMC vs CN and MCI



(c) LMCI vs. CN, MCI, Dementia folded trend chart



(d) EMC vs CN, MCI, Dementia fold trend graph

(e) Folding trend chart of AD vs Dementia

Figure 16 Fractal pattern of many different diseases over time

performed using SPSS with the above CN group as an example, as shown in the following figure fig. 17.

Using the above model to predict the y results for patients at the sample sites in 2022 is shown in Table.

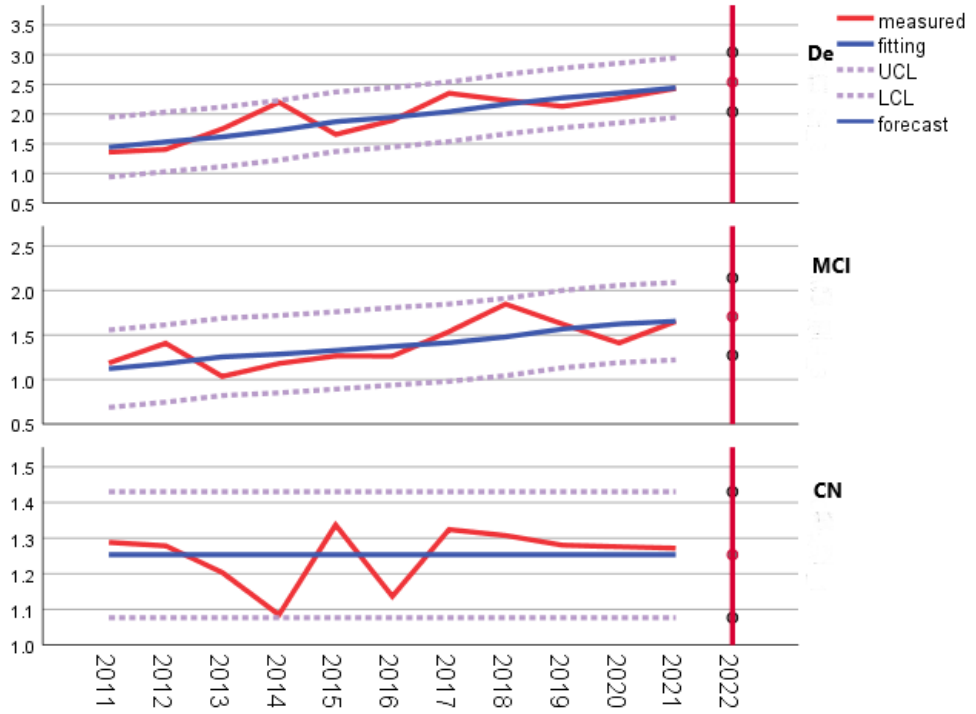


Figure 17 Time series prediction plots for the three scenarios of CN and final diagnosis

Table 5 Three Time Series Forecast Models on 2022 Forecast Results

	CN model	MCI model	De model
y predicted value	1.256	1.692	2.503

As shown above, for CN, the probability of not developing Dementia will remain the same within ten years, but the probability of developing MCI is higher than that of developing Dementia, and the trend of change is more stable without sudden change. For people with LMCI, only a small percentage of people will return to normal, most of them will deteriorate, and the condition of patients will deteriorate sharply within three years; for people with EMC, most of them will suffer from MCI, a small percentage will return to normal and suffer from Dementia, and the condition of patients will deteriorate within five years; for people with AD, there is a higher possibility of death and no sign of improvement within ten years.

4.5 Problem five modeling and solving

Our team has reviewed a large amount of literature and gives the following report. Early interventions and diagnostic criteria for different stages of the disease such as CN, SMC, EMCI, LMCI and AD are described in detail.

4.5.1 Early intervention

4.5.2 Early intervention in the pre-dementia stage of AD

①Early diagnosis and early intervention for preclinical AD.

②and control of risk factors in the pre-dementia stage of AD, combined with cognitive training has the potential to delay cognitive decline. For example, actively lowering lipids, regulating blood pressure, controlling blood sugar, improving mood, enhancing social interaction, ensuring good sleep, quitting smoking and alcohol, appropriate aerobic exercise, and timely detection and intervention of metal ion metabolism abnormalities. In addition, learn the Mediterranean dietary pattern, such as increasing vegetables, fruits, fish and beans and reducing red meat intake, and appropriately supplementing vitamins A, C, D, B6, B12 and folic acid. Chinese medicine also plays an important role in preventing the development of AD, such as acupuncture can improve the cognitive function of patients with mild cognitive impairment (MCI), and auricular acupressure therapy can stimulate the corresponding points to achieve the effect of health care and education. The regulation of intestinal flora has also become a new hot spot in AD prevention and treatment research.

③Early identification of the pre-dementia stage offers the possibility of early intervention in AD: Although we do not yet have drugs that can modify the disease process of AD, several international studies of early pharmacological interventions for the pre-dementia stage of AD have been initiated and are expected to find drugs that can modify the disease process of AD.

4.5.3 MCI phase intervention

MCI stage is the earliest stage of objective cognitive impairment in AD patients. Although MCI stage can also be divided into SMC, EMCI, and LMCI periods, there are only minor differences in patients' clinical symptoms and scale assessment status, so they can be grouped into MCI major categories to develop comprehensive intervention programs. At this stage, the patients' instrumental daily living ability is only slightly affected and their independence in life is still possible. Currently, the main MCI guidelines or expert consensus at home and abroad recommend a combination of non-pharmacological and pharmacological interventions for MCI.

Non-pharmacological treatment: It mainly includes moderate physical exercise, life behavior intervention, cognitive training, and socialization.

Pharmacological treatment: including ergot alkaloid preparations, ginkgo biloba extract, cholinesterase inhibitors, and glutamate receptor antagonists, etc. However, the efficacy of pharmacological treatment is limited in patients with amnestic MCI, and the pros and cons of pharmacological treatment of MCI are still debatable.

4.5.4 AD phase intervention

The 2018 Chinese guidelines for the diagnosis and treatment of dementia and cognitive impairment (II): guidelines for the diagnosis and treatment of Alzheimer's disease recommend.

Patients with a clear diagnosis of AD may be treated with the cholinesterase inhibitors ChEIs (Class A recommendation).

If treatment with a particular ChEIs is ineffective or is not tolerated due to adverse effects, the patient may be switched to another ChEIs or to a patch for treatment depending on the degree of adverse effects present at the patient's condition level (Grade B recommendation).

There is a dose-effect relationship, and high doses of ChEIs may be used as therapeutic agents in patients with moderate to severe AD, but the principle of gradual titration starting at low doses should be followed for administration (expert consensus).

Patients with clearly diagnosed moderate-to-severe AD can be treated with memantine or memantine in combination with donepezil and carboplatin, and the combination of ChEIs and memantine is particularly recommended for patients with severe AD who present with significant psycho-behavioral symptoms (Class A recommendation).

4.5.5 Diagnostic criteria

Currently, according to the international dementia diagnostic criteria Alzheimer's disease is generally classified into preclinical stage, mild cognitive impairment (MCI) and dementia (AD) and diagnostic criteria are developed as follows:

4.5.6 Pre -dementia stage

In 2011, the National Institute on Aging-Alzheimer's Association (NIA-AA) formally proposed a conceptual framework for preclinical AD. The criteria analyze preclinical AD, including stage -: stage of asymptomatic brain amyloidosis; stage 2: stage of amyloid-positive + synaptic dysfunction and/or early neurodegenerative changes, such as decreased tau protein in CSF, MRI suggestive of characteristic brain atrophy changes; stage 3: stage of amyloid-positive + evidence of neurodegeneration + very mild cognitive decline.

4.5.7 MCI

According to the 2018 Chinese Guidelines for the Diagnosis and Treatment of Dementia and Cognitive Impairment (V): Diagnosis and Treatment of Mild Cognitive Impairment, MCI is recommended to be diagnosed according to the above-mentioned international standards, based on several points: 1. objective evidence (from cognitive tests) of impairment in one or more domains of cognitive function; 2. complex instrumental daily abilities may be slightly impaired,

but independent daily living abilities are maintained, and the diagnosis of dementia has not been reached. A diagnosis of dementia has not been reached.

2: Etiological diagnosis: The etiological diagnosis of MCI was made by combining the onset and progression of the illness, the characteristics of the cognitive impairment, the history and signs of the presence or absence of a primary neurological disorder, psychiatric illness (or stressful event) or systemic disease, and the necessary ancillary tests.

4.5.8 AD

Diagnostic criteria specified by the National Institute of Neurological Disorders Speech-Language Disorders Stroke Association (NINCDS-ADRDA).

Core diagnostic criteria:

Early and significant situational memory impairment occurs, including the following features.

1. The patient or informant complains of slowly progressive memory loss for more than 6 months.

2. Tests reveal objective evidence of severe situational memory impairment: primarily impaired recall that does not significantly improve or return to normal by cueing or recognition testing.

3. At the onset or progression of AD, situational memory impairment may be independent of or associated with other cognitive function changes.

Criteria for confirming the diagnosis of AD.

The diagnosis of AD is confirmed if there is both clinical and histopathological (brain biopsy or autopsy) evidence, consistent with the NIA-Reagan criteria for autopsy confirmation of AD. There is both clinical and genetic (mutations on chromosomes 1, 14 or 21) evidence of AD diagnosis, and both criteria must be met.

4.5.9 SMC,EMCI,LMCI

However, for the three subdivided stages of the MCI process, SMC EMCI LMCI, the latest national and international guidelines for the diagnosis of dementia do not formally establish the corresponding diagnostic criteria. However, some studies have defined a simple diagnostic basis for it.

As in the INSIGHT-preAD study, the diagnostic criteria for SMC were.

”Are you complaining about your memory?” “Is it a regular complaint that has lasted now more than 6 months? ”

(1),the participating subjects answered ”yes” to both questions ”Are you complaining about your memory?” “Is it a regular complaint that has lasted now more than 6 months? ”.

(2),Subjects demonstrated intact cognitive functioning (i.e., meeting the objective requirement of no significant Alzheimer's disease) by three tests: the Brief Mental State Examination (≥ 27), the Clinical Dementia Rating Scale (0), and the FCSRT (Free and CuedSelective Rating Test) (total score ≥ 14).

In a study analyzing MRI analysis of Alzheimer's disease and the correlation between MRI and MMSE scores,

EMCI diagnostic criteria were: MMSE score of 24 to 30 (inclusive of 24 and 30) subjective memory events reported by the subject, reporter or clinician, objective memory deficits measured by education-adjusted scores, on delayed memory excerpted from the Wechsler Memory Scale Logical Memory 2 scale (WSLM2) (scored 9 to 11 for ≥16 years of education, 3 to 6 for 0-7 years scored 3-6) CRF of 0.5, no other significant cognitive impairment, basic maintenance of daily activities, and no dementia.

The diagnostic criteria for LMCI were the same as for EMCI, except that objective memory loss was measured according to adjusted education scores (<8 for those with ≥16 years of education, ≤4 for those with 8 to 15 years of education, and ≤2 for those with 0 to 7 years of education).

5. Strengths and Weakness

5.1 Advantages of the model

1. Question 1 used a number of statistical methods to analyze and test the data, on the basis of which the data results are more credible.
2. Question 2 uses multiple methods for comparison when exploring classification-only models, highlighting the goodness of the models in the paper.
3. The model for Problem 3 not only clusters the data set for visual presentation and presents its own analysis.
4. The model for question 4 uses stepwise regression to analyze the importance of time and indicators, and the selection of indicators is very reasonable and well done.

5.2 Disadvantages of the model

1. Data processing is relatively simple, and some important data may be lost.
2. More indicators can be considered to build an intelligent diagnostic model with multiple considerations.

References

[1] Qiu-Yang, Li-Sheng, Jin-Liang, Zhang-Mi-Mi, Wang-Jie. A method for identifying bridge anomaly monitoring data based on statistical feature mixing and random forest importance ranking[J]. Journal of Sensing Technology,2022,35(06):756-762.

[2] Sun Xiaochi. Machine learning methods in the diagnosis of Alzheimer's disease[D]. Zhongnan University of Economics and Law,2021,DOI:10.27660/d.cnki.gzczu.2021.002053.

[3] Huang, Invent, Chen, Jiawu, Fan, Xuanmei, Huang, Jinsong, Zhou, Chuangbing. Logistic regression fitting of temporal probability of rainfall-type landslides and continuous probability landslide hazard modeling[J/OL]. Earth Sciences:1-25[2022-11-21].http://kns.cnki.net/kcms/detail/42.1874.P.20211101.2007.018.html

[4] Sha Jinglan. A Comparative Study of P2P Internet Lending Default Prediction Models Based on LightGBM and XGBoost Algorithms [D]. Northeast University of Finance and Economics,2017

[5] Zhou LJ, Wang H, Wang WB, Zhang N. Parallel KMeans Algorithm for Massive Data[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition),2012,40(S1):150-152.DOI:10.13245/j.hust.2012.s1.022.

[6] shibing You. Yan Yan. Stepwise regression analysis method and its application[J]. Statistics and Decision Making,2017(14):31-35.DOI:10.13546/j.cnki.tjyjc.2017.14.007.

[7] Zhang, Sui, Zhang, Mei, Chi, Guotai. Entropy-based science and technology evaluation model and its empirical study[J]. Journal of Management,2010,7(01):34-42.

[8] Pan Difu, Liu Hui, Li Yanfei. Optimization model for wind speed prediction in wind farms based on time series analysis and Kalman filtering algorithm[J]. Grid Technology,2008(07):82-86.

Appendix

Listing 1: The first question python code

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
data = pd.read_excel("number.xlsx")
index = []
for column in list(data.columns):
    index.append(column)
    fig1 = plt.figure(1, facecolor='white', figsize=(10, 6))
    plt.rcParams['font.sans-serif'] = ['STKAITI']
    plt.rcParams['axes.unicode_minus'] = False
    plt.rcParams['axes.facecolor'] = '#cc00ff'
    plt.boxplot([data[column]],
                notch=True,
                sym='*',
                patch_artist=True,
                boxprops={'color': '#ffff00', 'facecolor': '#0066ff'},
                capprops={'color': '#ff3333', 'linewidth': 2},
                showmeans=True,
                meanline=True
                )
    plt.xticks(range(0, 1), [column], fontsize=20)
    plt.yticks(fontsize=20)
    plt.title(column, fontsize=25, color='#0033cc')
    plt.savefig("/{}.png".format(column),dpi=1000)
    plt.show()
a = pd.DataFrame()
b=0
print(index)

for i in index:
    Q1 = data[i].quantile(0.25)
    Q3 = data[i].quantile(0.75)
    IQR = Q3 - Q1
    print(IQR)
    print("", any(data[i] > Q3 + 1.5 * IQR))
    print(" ", any(data[i] < Q1 - 1.5 * IQR))
```

```
    fig2 = plt.figure(1, figsize=(10, 6))
    plt.style.use('ggplot')
    data[i].plot(kind='hist', bins=30, density=True)
    data[i].plot(kind='kde')
    plt.xlabel("x")
    plt.ylabel("y")
    plt.title(i)
    # print("")
    # plt.legend()
    plt.savefig("/{}.png".format(b),dpi=1000)
    plt.show()
    print("\n", data[i].describe())
    UB = Q3 + 1.5 * IQR
    LB = Q1 - 1.5 * IQR
    st = data[i].mean()
    print("", LB)
    print("", UB)
    print("", st)
    data.loc[data[i] > UB, i] = st
    data.loc[data[i] < LB, i] = st
    print("", data[i].describe())
    a.insert(b,i,data[i])
```

Listing 2: Second question python code

```
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
data=pd.read_csv(r"C:\Users\Administrator\Desktop\tomry\datazhong.csv",encoding='gbk')
data.columns
import numpy as np
import pandas as pd
from sklearn import preprocessing
X=data[[ 'Ventricles_bl', 'Hippocampus_bl', 'WholeBrain_bl',
       'Entorhinal_bl', 'Fusiform_bl', 'MidTemp_bl', 'ICV_bl', 'CDRSB_bl',
       'ADAS11_bl', 'ADAS13_bl', 'ADASQ4_bl', 'MMSE_bl', 'RAVLT_immediate_bl',
       'RAVLT_learning_bl', 'RAVLT_forgetting_bl', 'RAVLT_perc_forgetting_bl',
       'LDELTOTAL_BL', 'TRABSCOR_bl', 'FAQ_bl', 'MOCA_bl', 'EcogPtMem_bl',
       'EcogPtLang_bl', 'EcogPtVisspat_bl', 'EcogPtPlan_bl', 'EcogPtOrgan_bl',
       'EcogPtDivatt_bl', 'EcogPtTotal_bl', 'EcogSPMem_bl', 'EcogSPLang_bl',
```

```
      'EcogSPVisspat_bl', 'EcogSPPlan_bl', 'EcogSPOrgan_bl',
      'EcogSPDivatt_bl', 'EcogSPTotal_bl', 'ABETA_bl', 'mPACCdigit_bl',
      'mPACCtrailsB_bl']]
y=data['DX_bl']
import matplotlib.pyplot as plt
import numpy as np
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus']=False
get_ipython().run_line_magic('matplotlib', 'inline')
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import SelectFromModel
import numpy as np
import pandas as pd
from sklearn import preprocessing
X=data[['Ventricles_bl', 'Hippocampus_bl', 'WholeBrain_bl',
      'Entorhinal_bl', 'Fusiform_bl', 'MidTemp_bl', 'ICV_bl', 'CDRSB_bl',
      'ADAS11_bl', 'ADAS13_bl', 'ADASQ4_bl', 'MMSE_bl', 'RAVLT_immediate_bl',
      'RAVLT_learning_bl', 'RAVLT_forgetting_bl', 'RAVLT_perc_forgetting_bl',
      'LDELTOTAL_BL', 'TRABSCOR_bl', 'FAQ_bl', 'MOCA_bl', 'EcogPtMem_bl',
      'EcogPtLang_bl', 'EcogPtVisspat_bl', 'EcogPtPlan_bl', 'EcogPtOrgan_bl',
      'EcogPtDivatt_bl', 'EcogPtTotal_bl', 'EcogSPMem_bl', 'EcogSPLang_bl',
      'EcogSPVisspat_bl', 'EcogSPPlan_bl', 'EcogSPOrgan_bl',
      'EcogSPDivatt_bl', 'EcogSPTotal_bl', 'ABETA_bl', 'mPACCdigit_bl',
      'mPACCtrailsB_bl']]
y=data['DX_bl']
randomforest = RandomForestClassifier(random_state=0, n_jobs=-1)
model = randomforest.fit(X, y)
importances = model.feature_importances_
indices = np.argsort(importances)[::-1]
names = [X.columns[i] for i in indices]
print(names)
print(range(X.shape[1]), importances[indices])
plt.figure(figsize=(10, 7),dpi=80)
plt.suptitle('Importance of characteristics')
plt.bar(range(X.shape[1]), importances[indices],width=0.8)
plt.xticks(range(X.shape[1]), names, rotation=90)#rotation
plt.xlim([-1, X.shape[1]])
plt.tight_layout()
plt.savefig('')
```

```
plt.show()
sfm = SelectFromModel(randomforest, threshold=0.1, prefit=True) # prefit
X_selected = sfm.transform(X)
print('Number of features that meet this threshold criterion:',
     X_selected.shape[1])
from sklearn.model_selection import train_test_split
train_X,test_X,train_y,test_y =
   train_test_split(X,y,test_size=0.3,random_state=5)
from sklearn.metrics import classification_report,f1_score
import xgboost as xgb
import lightgbm as lgb
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score
import warnings
warnings.filterwarnings('ignore')
model_lr = LogisticRegression(
model_lr.fit(train_X,train_y)
print(classification_report(model_lr.predict(test_X),test_y))
model_lgb = lgb.LGBMClassifier()
model_lgb.fit(train_X,train_y)
print
print(classification_report(model_lgb.predict(test_X),test_y))
model_xgb = xgb.XGBClassifier()
model_xgb.fit(train_X,train_y)
print(classification_report(model_xgb.predict(test_X),test_y))
```

Listing 3: The third question python code

```
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.decomposition import PCA
import numpy as np
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
mapping2 = {
         'Male': 0,
         'Female': 1
```

```
}
mapping3 = {
          'White': 0,
          'Black': 1,
           'Asian':2,
           'More than one':3,
     'Unknown':4,
    'Am Indian/Alaskan':5,
    'Hawaiian/Other PI':6
}
mapping4 = {
          'Married': 0,
          'Widowed': 1,
          'Divorced': 2,
          'Never married':3,
          'Unknown':4
}
mapping5 = {
          'Not Hisp/Latino': 0,
          'Hisp/Latino': 1,
          'Unknown':2
}
data["DX_bl"] =data["DX_bl"].map(mapping)
data['PTGENDER'] =data['PTGENDER'].map(mapping2)
data['PTRACCAT'] =data['PTRACCAT'].map(mapping3)
data['PTMARRY'] =data['PTMARRY'].map(mapping4)
data['PTETHCAT'] =data['PTETHCAT'].map(mapping5)
data4=data[data['DX_bl']==2]
data4=data4.drop('DX_bl',axis=1)
print(data4)
select_cols = data4.columns
gaojia_df = data4
km = []
sses = []
for k in range(2, 10):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(gaojia_df[select_cols])
    sses.append(kmeans.inertia_)
    km.append(k)
```

```
plt.figure(figsize=(3, 2), dpi=300)
plt.tick_params(labelsize=4)
plt.plot(km, sses, color='coral')
plt.show()
df1=data4
data1 = np.array(df1)
clf1 = AgglomerativeClustering(n_clusters = 3, linkage = 'ward')
s = clf1.fit(data1)
pred1 = clf1.fit_predict(data1)
score1 = silhouette_score(data1, pred1)
pca = PCA(n_components=2)
newData1 = pca.fit_transform(data1)

x1, y1 = [], []
x2, y2= [], []
x3, y3= [
```

Listing 4: Fourth question matlab code

```
load matlab.mat
x=2011:1:2021;
x1=2013:1:2021;
x2=2011:1:2019;
x3=2012:1:2021;
figure(1)
plot(x,y1,'-Xg',x,y2,'-*r',x,y3,'-ob');
 title(['CN trilinear chart'])
legend('CN-CN','CN-Dementia','DN-MCI');
xlabel('year')
ylabel('ecog composite index')
figure(2)
plot(x1,y4,'-Xg',x1,y5,'-*r');
 title(['SMC double line diagram'])
legend('SMC-CN','SMC-MIC');
xlabel('year')
ylabel('ecog composite index')
figure(3)
plot(x3,y6,'-Xg',x3,y7,'-ob',x2,y8,'-*r');
 title(['LMCI trilinear chart'])
legend('LMCI-CN','LMCI-MCI','LMCI-Dementia');
```

```
xlabel('year')
ylabel('ecog composite index')
figure(4)
plot(x2,y9,'-Xg',x3,y10,'-ob',x,y11,'-*r');
title(['EMCI trilinear chart'])
legend('EMCI-CN','EMCI-MCI','EMCI-Dementia');
xlabel('year')
ylabel('ecog composite index')
figure(5)
plot(x3,y12,'-*r');
 title(['AD Line Chart'])
legend('AD-Dementia');
xlabel('year')
ylabel('ecog composite index')
```