| C | 2022 ShuWei Cup Summary | Team Control Number 2022111614350 |
|---|---|---|

Alzheimer's disease (AD) is a chronic progressive neurodegenerative disease。This paper explores the problem of classifying AD categories and establishing diagnostic criteria by using random forest algorithm, XGBoost algorithm, support vector machine, and time series analysis to build models.

**Task1:** Firstly,the data is cleaned using python for the AD feature variables given in the attachment. The random forest model was then applied to generate variable importance plots to obtain the strength of correlation between the characteristic variables and Alzheimer's disease diagnosis, and four characteristic variables, LDELTOTAL, IMAGEUID, mPACCtrailsB, mPACCdigit, and PTEDUCAT, were obtained to have strong correlation with AD, and then five types of box line plots were drawn about these four characteristic variables with box line plots, the results are maintained at the same level and the results of random forest are verified.

**Task2:** This paper uses the mainstream classification algorithms in machine learning (random forest, XGBoost and SVM) to build three models based on structural brain features and cognitive behavioral feature variables respectively to study the criteria for diagnosing AD, divides the dataset into two categories, AD and non-AD, and repeatedly trains the model until it can achieve predetermined accuracy on the training dataset. The accuracy of the training set, cross-validation set and test set of the random forest model is 0.896, 0.824 and 0.841, respectively, which is better than that of XGBoost and SVM than the random forest model.

**Task3:** The random forest and decision tree methods with better results in problem 2 were chosen to classify the data set. The data were firstly processed into three major categories of CN, MCI and AD for classification model building, and then three minor categories of SMC, EMCI and LMCI were extracted for classification model building to achieve the purpose of cluster classification and analyzed to get the accuracy of training set, cross-validation set and test set all exceeded 80%, and the classification effect was significant.

**Task4:** The highly correlated characteristic variables LDELTOTAL and mPACCDigit from question 1 were selected, and time series models ARIMA model (0,1,0) and ARIMA model (0,0,2) were developed to explore the patterns of different categories of disease evolution over time. high values of LDELTOEAL and mPACCdigit affect Alzheimer's disease type diagnosis.

**Task5:** This paper cites a large number of authoritative journals and literature on AD, specifically analyzes the development history and details of the five types of diagnostic criteria of CN, SMC, EMCI, LMCI and AD, and gives suggestions for the diagnostic criteria of AD in conjunction with this paper to provide feasible solutions for the diagnostic criteria of AD.

***Keywords :Alzheimer's disease; random forest algorithm; XGBoost; SVM; decision tree; time series model***

# Content

# 1. Restatement of the problem

Alzheimer's disease (AD) is a neurodegenerative disease that develops in the elderly population and is characterized by cognitive and intellectual impairment, as well as a decline in behavioral and life skills. The disease process of AD can be divided into three stages: prodromal AD, Mild Cognitive Impairment (MCI) and AD. MCI is a transitional state between normal aging and dementia, and if treated thoroughly in the early stages of AD, the onset of AD can be greatly delayed.

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

Question 2: Using additional structural brain features and cognitive-behavioral features to design an intelligent diagnosis of Alzheimer's disease

Question 3: First, CN, MCI and AD are divided into three major categories. Then, for the three subclasses contained in MCI (SMC, EMCI and LMCI),the cluster continues to be refined into three subclasses.

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

Question 5: Please refer to 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 Problem 1

Question 1 requires us to preprocess the characteristic variables of the data and to investigate the correlation between the data characteristics and the diagnosis of Alzheimer's disease. The characteristic variables in the data must first be understood by reviewing relevant information, and then the raw data set must be observed and analyzed, thus revealing a large number of missing values and variables with overlapping information in the data. Therefore, data processing software such as python is needed to clean the original data. Specifically, the characteristic variables with less correlation, overlapping information, and more missing values need to be removed, and the remaining small number of missing values need to be filled using the mean fill method, and finally the qualitative variables are converted into quantitative variables. For the correlation between data features and Alzheimer's disease diagnosis, the random forest model was mainly used to generate variable importance graphs to obtain the strength of correlation between feature variables and Alzheimer's disease diagnosis, and then the results obtained from the random forest were verified by box plots of feature variables with respect to the type of Alzheimer's

disease diagnosis.

## 2.2 Analysis of Problem 2

Problem 2 requires us to design an Alzheimer's disease diagnosis using additional structural brain features and cognitive-behavioral features. We classified DX_bl diagnoses into two categories, one normal and the second diseased, by comparing several mainstream classification algorithms (Random Forest, XGBoost, SVM) currently used in supervised learning for ADNIMERGE_NewDataset classification was performed. A model is built using structural brain features and cognitive-behavioral feature variables so that for known values of the predictor variables, the corresponding values of the target variables are obtained. This model is repeatedly trained until it can achieve a predetermined accuracy on the training dataset. The fit of the different models is then compared and the optimal model is selected to diagnose Alzheimer's disease.

## 2.3 Analysis of Problem 3

Problem 3 requires us to first classify CN, CMI and AD into three major classes, and then for the three subclasses contained in MCI (SMC, EMCI and LMCI), the clusters continue to be refined into three subclasses. To complete the classification of the data, the random forest and decision tree methods were chosen to be used for the classification. The data were first processed into three major classes for classification model building, and then the three subclasses (SMC, EMCI and LMCI) were extracted for classification model building to achieve the purpose of cluster classification and analyze its effect.

## 2.4 Analysis of Problem 4

Question 4 asks us to analyze the same sample in the annexes based on features containing collections at different time points, in conjunction with time points, in order to discover patterns in the evolution of different categories of diseases over time. For this purpose, subjects with data from multiple different time points were first selected, then those with the same number of collections were filtered and those containing a large number of missing values were excluded. The experimental data obtained were used to select the important characteristic variables to be studied. Since there are differences in the data of individual subjects, we averaged multiple subjects for the same index to eliminate their individual differences and thus to study their evolution patterns over time.

## 2.5 Analysis of Problem 5

Question 5 asked us to refer to the relevant literature to describe the early intervention and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI, and AD,. To address this goal, this paper read several authoritative literature and journals to integrate the history of AD diagnostic criteria so as to obtain the development of AD diagnostic criteria and relate them to this paper, hoping to contribute some new ideas to AD diagnostic criteria.

# 3.Model Assumptions

Hypothesis 1: The data for the Alzheimer's disease characteristic variables in Annex 1 are really reliable.
Reason 1: The authenticity of the data can be applied to the solution and validation of the model, otherwise the authenticity of the data is questionable, and the obtained model results will have errors and cannot truly reflect the relationship between variables.
Hypothesis 2: Alzheimer's disease is disturbed only by structural brain features and cognitive-behavioral features and is not influenced by other features.
Reason 2: Alzheimer's disease is also influenced by age and years of education, as well as other characteristics, which are not considered in this paper.
Hypothesis 3: The classification of each population is a reflection of the influence of the characteristic data, and there is no case that the characteristic data are wrong for the disease.
Reason 3: The classification of each population is a reflection of the impact of the feature data, in order to meet the reliability of the subsequent modeling classification.
Hypothesis 4: It is assumed that the two data selected in this paper are most strongly correlated with the disease and that the other characteristic data correlations do not affect the development of the disease.
Reason 4: Other feature data in the evolution over time also generate to the disease evolution over time, and the case is not considered here in this paper.

# 4.Definition and Symbol Description

| AD | Alzheimer's disease |
|---|---|
| $Y_t$ | $t$ Sequence value at the moment |
| $\theta(B)$ | Moving average coefficient polynomial |
| $\Phi(B)$ | Autoregressive coefficient polynomial |

| | |
|---|---|
| $\nabla^d$ | $d$ Step Difference |
| $\varepsilon_t$ | $t$ Residuals at the moment |
| $h(x_i)$ | Hypothesis function for the $i$ data |
| $y_i$ | Category of the ith data |
| $w_k$ | Weight coefficients for linear weighted regression |
| $d$ | Delineate the distance between the hyperplane and the sample |
| $k(x_i, x_j)$ | First $i$ data to the first $j$ data |

*Other unmarked symbols are indicated in the text

# 5.Model building and solving

## 5.1 Question 1

Random Forest (RF) is an extended variant of Bagging. RF further introduces random attribute selection in the training process of decision tree based on the decision tree as the base learner to build Bagging integration. Specifically, the traditional decision tree selects an optimal attribute in the attribute set (assumed to have d attributes) of the current node for division; while in RF, for each node of the base decision tree, a subset containing k attributes is first randomly selected from the attribute set of that node, and then an optimal attribute is selected from this self for division. Here the parameter k controls the degree of randomness invoked: if k = d, the base decision tree is constructed the same as the traditional decision tree; if k = 1 is made, then it is a random selection of an attribute for division; in general, the recommended value k = log2d.

### 5.1.1 Data pre-processing

In ADNIMERGE_New, we found a large number of missing values and complex variables in the dataset. In response to this situation, we first conducted a preliminary screening of the data, and the process is as follows.

Based on the understanding of each variable and the purpose of the study, we removed the historical data recorded by the first time the investigator received a data collection, i.e., the variables containing the "b1" suffix, and only intercepted the current value of a data item for the study. The remaining 64 variables. Then the

characteristic variables with less correlation or overlapping information of DK categorical variables were deleted, such as: COLPROT, ORIGPROT, PTID, SITE, VISCODE, EXAMDATE, FLDSTRENG, FSVERSION, Month, EcogPtMem, EcogPtLang, EcogPtVisspat, EcogPtPlan, EcogPtOrgan, EcogPtDivatt, EcogPtTotal, EcogSPMem, EcogSPLang, EcogSPVisspat, EcogSPPlan, EcogSPOrgan, EcogSPDivatt, EcogSPTotal, and 30 feature variables were removed. Finally, the feature variables with a large degree of missing feature variables were deleted, because the degree of missing feature variables is too large is not conducive to data analysis and research, even if various methods of dealing with missing values are used to make up for the missing values, that greatly destroys the authenticity of the data, and the analysis results obtained are no longer convincing. After the final data screening we obtained a total of 24 feature variables for the study, the variables are as follows: AGE, PTGENDER, PTEDUCAT, PTRACCAT, PTETHCAT, PTMARRY, ADAS11, ADAS13, ADASQ4, MMSE, RAVLT_immediate, RAVLT_learning, RAVLT_forgetting, RAVLT_perc_forgetting, LDELTOTAL, TRABSCOR, FAQ, mPACCdigit, mPACCtrailsB, IMAGEUID, Ventricles, WholeBrain, ICV, and DX.

For some of the filtered feature variables there are still missing data, so there is a need to do further filling process for these 24 variables. In this problem, the mean value approach is used to find the mean value of the characteristic variables with missing values by category (CN, AD, LMCI, EMCI, SMC), and then the mean value is inserted into the corresponding missing values.

Since there are also qualitative variables in the characteristic variables, it is also necessary to convert the qualitative variables to a variable form similar to (0,1). The variables are converted as follows.

Table 5-1.Variable Conversion

| PTRACCAT | DX | PTMARRY | PTETHCAT | PTGENDER |
|---|---|---|---|---|
| White:0 | CN: 0 | Never married:0 | Not Hisp/Latino:0 | Male: 0 |
| Black:1 | AD:1 | Married:1 | Hisp/Latino:1 | Female:1 |
| Asian:2 | LMCI:2 | Divorced:2 | Unknown:2 | |
| More than one:3 | EMCI:3 | Widowed:3 | | |
| Unknown:4 | SMCI:4 | Unknown:4 | | |
| Am Indian/Alaskan:5 | | | | |
| Hawaiian/Other PI:6 | | | | |

After the above process we finally finished processing the data.

## 5.1.2 Exploring the relevance of data features to the intelligent diagnosis of Alheimer's disease.

### 5.1.2.1 Completion of importance assessment method using random forest model

Random Forest (RF) is an extended variant of Bagging. Based on the decision tree-based learner to build the Bagging ensemble, RF further introduces random attribute selection in the training process of decision tree. Specifically, in a random forest, for each node based on the decision tree, a subset containing k attributes is randomly selected from the set of attributes of the node, and then an optimal attribute is selected from this self for division.

The random forest model will be used in this problem to explore the correlation between data features and intelligent diagnosis of Alheimer's disease. In fitting the random forest model, it can give the importance of each variable in the classification and thus reflect the magnitude of the correlation between the feature variables and the response variables. The following graphs of the importance of the response variables were obtained from the developed random forest model.

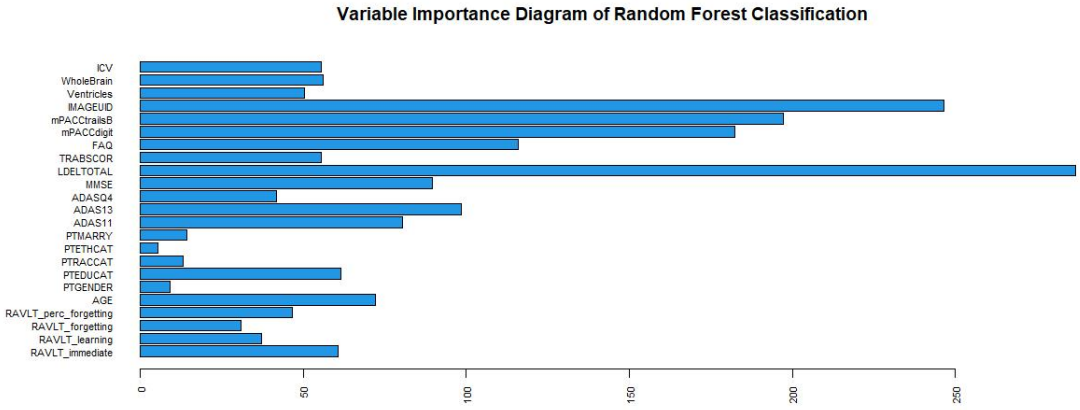


Figure: 5-1 Variable importance map of random forest classification for Alzheimer's disease data

Using the above random forest plot of variable importance we found that the variables LDELTOTAL, IMAGEUID, mPACCtrailsB, and mPACCdigit had a decisive role in making Alheimer's disease classification, while the variables PTGENDER, PTRACCAT, PTETHCAT, and PTMARRY had relatively the diagnosis of Alzheimer's disease was less influential.

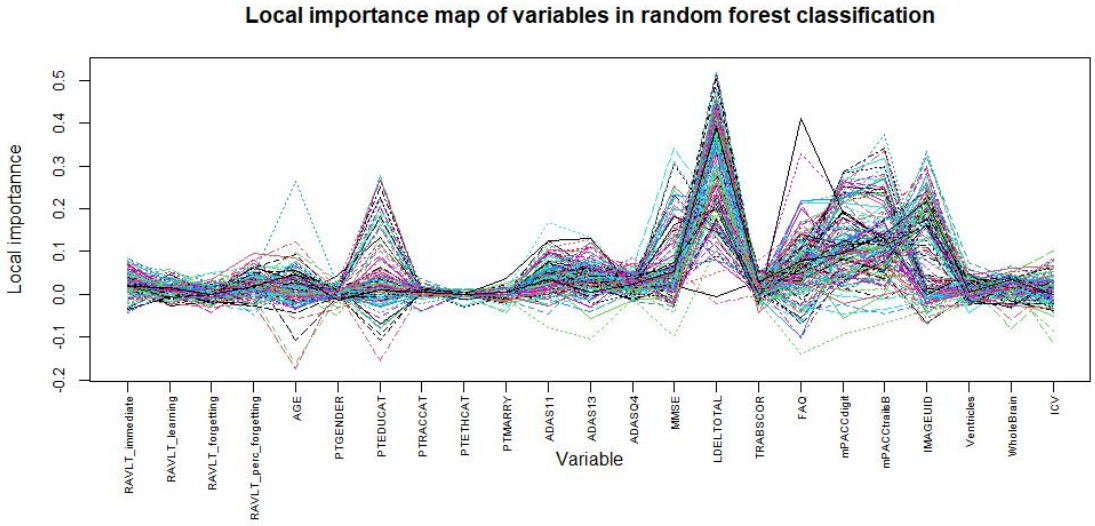


Figure: 5-2 Variable Local Importance Map for Random Forest Classification of Alzheimer's Disease Data

The above figure represents the local importance of the 200 randomly selected observations for each feature variable, from which it can be seen that the feature variables LDELTOTAL, IMAGEUID, mPACCtrailsB, mPACCdigit, and PTEDUCAT are relatively more important, and the results obtained are consistent with those obtained in Figure 5-1.

### 5.1.2.2 Analysis of correlations between characteristic variables and Alzheimer's disease diagnosis using box plots

Combining the magnitude of the correlation between the feature variables obtained from the random forest and the intelligent diagnosis of Alheimer's disease, several of them were selected to verify whether the change was relevant.

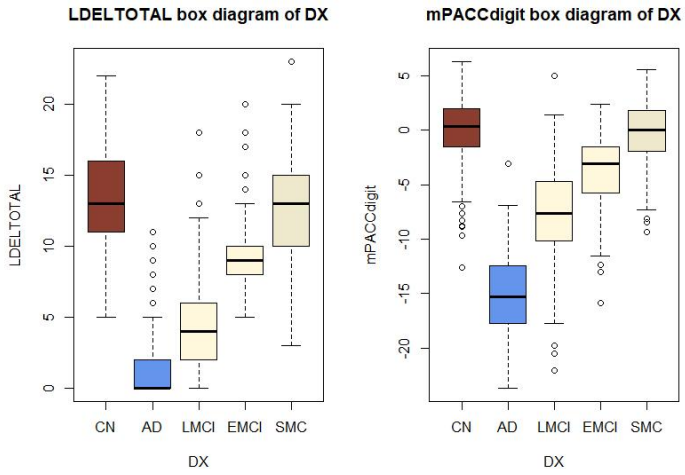


Figure 5-3 LDELOTAL and mPACCDigit box diagram of DX

In the above figure, two characteristic variables of greater importance, LDELTOTAL and mPACCdigit, were selected and their box plots regarding the

diagnostic categories (CN, AD, LMCI, EMCI, SMC) were drawn separately. From the left panel, it can be found that for normal elderly people (CN) the LDELOTAL values are relatively concentrated around 13, for Alzheimer's disease patients (AD) the LDELOTAL values are relatively concentrated around 1, while MCI patients (LMCI, EMCI, SMC) are distributed between CN and AD in order of severity depending on the disease. From the right panel, we can find that the mPACCDigit values are relatively concentrated around 0 for normal elderly (CN), around -15 for Alzheimer's disease patients (AD), and between CN and AD for MCI patients (LMCI, EMCI, SMC) depending on the severity of the disease. Through the analysis, we found that LDELTOTAL and mPACCdigit had a great correlation with the diagnosis of Alzheimer's disease, and the magnitude of LDELTOTAL and mPACCdigit values largely influenced the diagnosis of Alzheimer's disease.

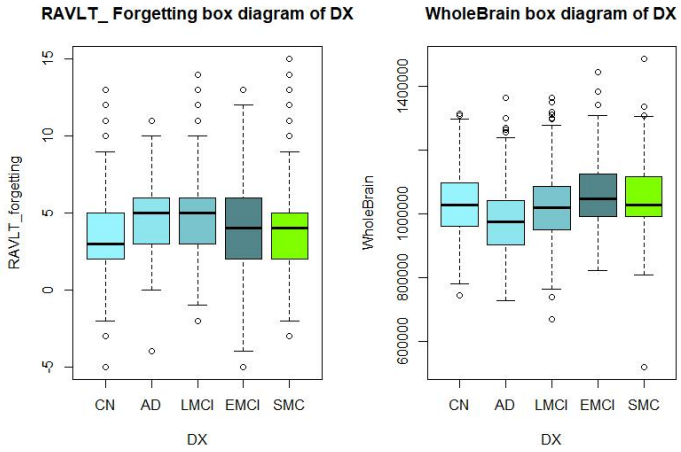


Figure 5-4 RAVLT_ Forgetting and WholeBrain box diagram of DX

The above figure selects two less important two characteristic variables RAVLT_ Forgetting and WholeBrain, and draws their box plots about the diagnostic categories (CN, AD, LMCI, EMCI, SMC) respectively. Both the left and right plots we can find that for different populations RAVLT_ Forgetting and WholeBrain always maintain almost the same level, with RAVLT_ Forgetting concentrated roughly in the interval [3, 5] and WholeBrain, roughly in the interval [90,000, 110,000]. Therefore, it can be concluded that these two characteristic variables alone cannot be well applied to the diagnosis of Alheimer's disease, indicating that the correlation between these two variables is low for the diagnosis of Alheimer's disease.

## 5.2 Question 2.

### 5.2.1 Building Random Forest, XGBoost and SVM Models

In Problem 1, we have used the random forest method to obtain that some of the data features are indeed correlated with Alzheimer's disease. In Problem 2, we will use the feature variables obtained in Problem 1 to build a new model without going

into too much detail about the random forest. The XGBoost and SVM models used are highlighted below

#### 5.2.1.1 XGBoost

XGBoost (eXtreme Gradient Boosting), is an algorithm or engineering implementation based on GBDT. Gradient Boosting Decision Tree (GBDT) is an additive model based on the idea of boosting integration, where the training is done greedily using a forward distribution algorithm, and a CART tree is learned at each iteration to fit the residuals of the prediction results of the previous t-1 trees to the true values of the training samples. The objective function of XGBoost consists of two parts: loss function and regularization.

Derivation of the objective function of XGBoost: the training dataset $T = \{(\mathrm{x1}, y1), (x2.y2), ..., (xn, yn)\}$ is known, and the loss function :

$$l(yi, \hat{y}i)$$

Regularization term:

$$\Omega(f_k)$$

Then the overall objective function can be written as

XGBoost with GBDT gradient boosting tree expression.

Due to

$$\hat{y}_i = \sum_{k=i}^{k} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Translate the objective function into.

$$\Im^{(t)} = \sum_{\mathrm{i}}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(xi)) + \sum_{k} \Omega(f_k)$$

Next, three steps optimize the XGBoost objective function

Step 1: second-order Taylor expansion, removing the constant term and optimizing the loss function term.

Step 2: regularization expansion, removal of constant terms and optimization of regularization terms.

Step 3: Combine the primary term coefficients and quadratic term coefficients to obtain the final objective function

Definition.

$$G_{\mathrm{j}}=\sum_{\mathrm{i}\in I_{\mathrm{j}}}\mathrm{g_i}\ H_{\mathrm{j}}=\sum_{\mathrm{i}\in I_{\mathrm{j}}}\mathrm{h_i}$$

Bringing the above into the objective function yields.

$$\mathfrak{I}^{(\mathrm{t})}=\sum_{\mathrm{j}=1}^{T}[G_j w_j+\frac{1}{2}(H_j+\lambda){w_j}^2]+\gamma T$$

The final objective function of XGBoost is obtained.

The objective function is known, then the objective function of each leaf node is

$$f(w_j)=G_j w_j+\frac{1}{2}(H_j+\lambda){w_j}^2$$

Its a quadratic function.

$$(H_j+\lambda)>0, and f(w_j) in W_j=-\frac{G_j}{H_j+\lambda}\ get\min inum=-\frac{1}{2}\frac{{G_j}^2}{H_j+\lambda}$$

If the target value Obj is the smallest, the tree structure is the best, and this is the optimal solution of the objective function

The objective equations of each leaf node of the XGBoost objective function are independent of each other.

That is, each leaf system hey but the equation reaches the maximum value point and the whole objective function reaches the maximum value point.

Then the weight wj of each leaf node and the objective value of Obj that reaches the optimum at this time.

$$w_j=-\frac{G_j}{H_j+\lambda}\ Obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{{G_j}^2}{H_j+\lambda}+\gamma T$$

The objective Obj value is the smallest, then the tree structure is the best, which is the optimal solution of the objective function at this time.

### 5.2.1.2 SVM

Support vector machines (SVMs) are a binary classification model whose basic model is a linear classifier defined by interval maximization on the feature space, and interval maximization distinguishes it from a perceptron; SVMs also include kernel tricks, which make them substantially nonlinear classifiers. The learning algorithm of SVM is the optimization algorithm for solving convex quadratic programming.

The basic idea of SVM learning is to solve the separated hyperplane that can

correctly partition the training dataset with maximum geometric interval. sVM has the ability to find the optimal solution based on finite samples and can avoid the problem of avoiding local extrema in neural networks to obtain the global optimal point and high-dimensional feature processing capability. In this paper, support vector machine is chosen as a classifier to distinguish AD from non-AD. Finally, the hyperplane with the best classification parameters is used to discriminate the feature vectors in the text to be measured.

## 5.2.2 Design and implementation of supervised classification model

### 5.2.2.1 Random Forest Model

For the diagnosis of Alzheimer's disease, we used structural brain characteristics and cognitive-behavioral characteristics as the conditions for the determination of Alzheimer's disease. Due to the large missing values of some variables, we screened the variables and finally selected some parameters of brain characteristics and cognitive-behavioral characteristics: 'MMSE', 'RAVLT_immediate', 'RAVLT_learning', ' RAVLT_forgetting', 'RAVLT_perc_forgetting', 'LDELTOTAL', 'FAQ', 'Ventricles', 'WholeBrain', 'ICV' were used as fitting variables of the diagnostic model for model fitting.

Then the encoding is mapped and coded, the gini node splitting evaluation criterion is selected, the amount of control decision tree is 100, the maximum depth of the tree is 10, the random forest model is established to evaluate the selected variables for refitting, and the accuracy of the test set and training set is output to judge the goodness of the model fit.

**5.2.2.2 XGBOOST model**

For features that have been scaled, normalized and used as classifiers . Use default parameters for xgboost-XGBClassifier and use a tree based model for computation. Fit the classifier model with the trainer data, then fit the classifier model with the trainer data X and make predictions on the trainer data X. Output the prediction results and show a text report of the classification metrics. The performance of the XGBoost model is evaluated by k-fold cross-validation using the training and test sets, and the cross-validation method is used to evaluate the model performance with less variance than the method using a single training and test set split. It does this by splitting the dataset into k partial datasets each split is called a fold and the algorithm trains on the k-1 fold and tests on the remaining one, by repeating so that each fold has a chance to be used as a test set. After running cross-validation, k different performance scores are obtained, and the mean and standard deviation of these score scores are obtained to measure the performance of the model. Evaluation of the algorithm by this method is more reliable because the algorithm is trained and evaluated several times on different data remember.

### 5.2.2.3 SVM model

The processed data is parametrized into GridSearchCV, the classifier model is fitted with the trainer data using the svc model, the classifier model is fitted with the trainer data X and the prediction is performed on the trainer data X. The optimal result is output 100 iterations.

## 5.2.3 Presentation of results.

Table 5-2. Evaluation results of random forest model

| | Accuracy | Recall Rate | Accuracy rate | F1 |
|---|---|---|---|---|
| Training set | 0.896 | 0.896 | 0.893 | 0.894 |
| Cross-validation sets | 0.824 | 0.824 | 0.823 | 0.821 |
| Test set | 0.841 | 0.841 | 0.842 | 0.841 |

Table 5-3.XGBOOST model evaluation results

| | Accuracy | Recall Rate | Accuracy rate | F1 |
|---|---|---|---|---|
| Training set | 0.863 | 0.863 | 0.864 | 0.864 |
| Cross-validation sets | 0.831 | 0.831 | 0.832 | 0.831 |
| Test set | 0.831 | 0.831 | 0.841 | 0.835 |

Table 5-4. SVM model evaluation results

| | Accuracy | Recall Rate | Accuracy rate | F1 |
|---|---|---|---|---|
| Training set | 0.681 | 0.681 | 0.64 | 0.658 |
| Cross-validation sets | 0.69 | 0.69 | 0.636 | 0.658 |
| Test set | 0.68 | 0.68 | 0.649 | 0.663 |

We use 80% of the original data as theTraining set and 20% as the test set. The original training set (distinguished from the test set data) was first divided into 5 parts, and then 4 of the data were selected for training and the remaining 1 part was used to evaluate the effect, which was repeated 5 times to ensure that each part of the data was used as over training data and validation data. From the table of evaluation results of different models, it can be seen that the highest accuracy of the training set and test set for Alzheimer's disease is the random forest model. Therefore, in designing the Alzheimer's disease diagnosis building model using structural brain features and cognitive behavioral features, we decided to use the random forest model.

## 5.3 Question 3

### 5.3.1 Classifying CN, MCI and AD

Based on the 24 random variables obtained from question one, clusters were first classified for the three major categories of CN, MCI (SMC, EMCI, LMCI) and AD. Combined with the importance plot of the characteristic variables of question one, it was found that there existed some variables (PTMARRY, PTETHCAT, PTRACCAT, PTGENDER) that were less relevant for the diagnostic classification of Alzheimer's disease and less important for the cluster classification of question three, so this part was chosen to be excluded. The remaining 19 characteristic variables and one response variable were used for cluster classification of the three major categories of CN, MCI, and AD.

#### 5.3.1.1 Exploring the correlation between characteristic variables

The correlation coefficient matrix between the 19 characteristic variables is first calculated, and then the correlation heat map between the 19 characteristic variables is output by visualization method, and the results are as follows.

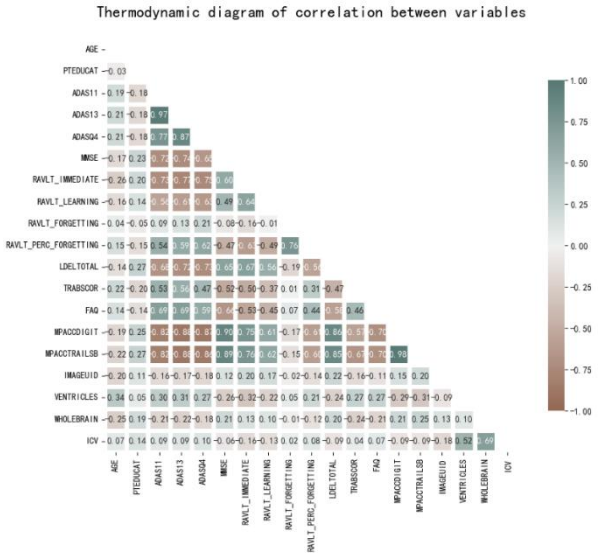


Figure 5-5.Thermodynamic diagram of correlation between variables

The correlation graph of feature variables obtained from the output shows that there are correlations among 19 feature variables, so the classification method of plain Bayes cannot be used to classify the data into three major categories, CN, MCI and AD, by using decision tree and random forest methods in the following.

### 5.3.2 Classification of CN, MCI and AD using Random Forest and Decision Trees

The data clusters were classified into 3 broad categories based on 19 feature variables, and a 5-fold cross-validation method was used to evaluate the classification accuracy and compare the excellence of different models. In which 2425 experimental data were intercepted and both models were randomly selected with 70% of the data as training set and 30% of the data as test set, the confusion matrix and model evaluation results are as follows.

Table 5-5.Decision tree model evaluation results

| | Accuracy | Recall Rate | Accuracy rate | F1 |
|---|---|---|---|---|
| **Training set** | 0.925 | 0.925 | 0.925 | 0.925 |
| **Cross-validation sets** | 0.811 | 0.811 | 0.813 | 0.81 |
| **Test set** | 0.85 | 0.85 | 0.849 | 0.849 |

Table 5-6.Confusion matrix of decision tree predictions and true categories

| | | Real categories | | |
|---|---|---|---|---|
| | | CN | AD | MCI |
| Categories of predictions | CN | 120 | 0 | 37 |
| | AD | 0 | 92 | 27 |
| | MCI | 27 | 18 | 407 |

From the above two obtained are the evaluation results and confusion matrix using the decision tree classification model. From the evaluation results, we can see that the accuracy of the model for the training set is 92.5%, which proves that the fit of the model is good; and the accuracy for the cross-validation set and the test set is 81.1% and 85%, respectively, which proves that the generalization ability of the model can also be guaranteed. The diagonal elements in the confusion matrix are the number of people who are correctly classified, while the elements on the non-diagonal line are the number of people who are incorrectly classified. From the table, we can find that the accuracy of predicting CN is 81.6%, the accuracy of predicting AD is 83.6%, and the accuracy of predicting MCI is 86.4%, and the accuracy of predicting MCI is higher in comparison.

Table 5-7.Random Forest Model Evaluation Results

| | Accuracy | Recall Rate | Accuracy rate | F1 |
|---|---|---|---|---|
| **Training set** | 0.905 | 0.905 | 0.911 | 0.903 |
| **Cross-validation sets** | 0.844 | 0.844 | 0.849 | 0.839 |
| **Test set** | 0.842 | 0.842 | 0.85 | 0.835 |

Table 5-8.Random Forest Confusion matrix of predicted and true categories

| | Real categories |
|---|---|

| | | CN | AD | MCI |
|---|---|---|---|---|
| Categories of predictions | CN | 96 | 0 | 69 |
| | AD | 0 | 109 | 24 |
| | MCI | 11 | 11 | 408 |

The two obtained above are the evaluation results and confusion matrix using the random forest classification model, from which it can be seen that the accuracy of the model for the training set is 90.5%, which is slightly lower than that of the decision tree model in comparison; the accuracy of the cross-validation set is higher than that of the decision tree and the accuracy of the test set is slightly lower than that of the decision tree, with results of 84.4% and 84.2%, respectively. The accuracy of predicting CN in the confusion matrix was obtained as 89.7%, predicting AD as 90.8%%, and predicting MCI as 81.4%.

Both the decision tree model and the random forest model are good for classifying CN, MCI, and AD categories, but each has its own focus. In contrast decision trees are more accurate for predicting MCI, while random forests are more accurate for predicting CN and AD.

### 5.3.3 Classifying the three subclasses of MCI (SMC, EMCI and LMCI)

Combining the method of the previous section to continue the classification using random forest and decision tree, the same first removed four variables with less importance (PTMARRY, PTETHCAT, PTRACCAT, PTGENDER). Then the samples belonging to the CN and AD categories were removed from the data, leaving only three categories, SMC, EMCI and LMCI, and finally 1473 experimental data were obtained. The results of the analysis were as follows.

Table 5-9.Decision tree model evaluation results

| | **Accuracy** | **Recall Rate** | **Accuracy rate** | **F1** |
|---|---|---|---|---|
| **Training set** | 0.987 | 0.987 | 0.987 | 0.987 |
| **Cross-validation sets** | 0.891 | 0.891 | 0.894 | 0.891 |
| **Test set** | 0.88 | 0.88 | 0.881 | 0.881 |

Table 5-10.Confusion matrix of decision tree predictions and true categories

| | | Real Categories | | |
|---|---|---|---|---|
| | | LMCI | EMCI | SMC |
| Categories of predictions | LMCI | 200 | 4 | 4 |
| | EMCI | 2 | 105 | 19 |
| | SMC | 3 | 21 | 84 |

From the evaluation results and the confusion matrix obtained above, it can be seen that the accuracy of the model for the training set is as high as 98.7%, which proves that the model is a very good fit, but there is a risk of overfitting; however, the

accuracy for the cross-validation and test sets is also as high as 89.1% and 88%, which proves that the generalization ability of the model is also quite good without overfitting. From the confusion matrix, we can see that the accuracy of predicting LMCI is 97,6%, the accuracy of predicting EMCI is 80.8%, and the accuracy of predicting SMC is 78.5%. The accuracy of the decision tree for predicting LMCI is quite high, but EMCI and SMC are a little bit lower.

Table 5-11.Random Forest Model Evaluation Results

| | **Accuracy** | **Recall Rate** | **Accuracy rate** | **F1** |
|---|---|---|---|---|
| **Training set** | 0.969 | 0.969 | 0.97 | 0.969 |
| **Cross-validation sets** | 0.908 | 0.908 | 0.908 | 0.907 |
| **Test set** | 0.912 | 0.912 | 0.911 | 0.912 |

Table 5-12.Confusion matrix of random forest predictions and true categories

| | | Real Categories | | |
|---|---|---|---|---|
| | | LMCI | EMCI | SMC |
| Categories of predictions | LMCI | 201 | 4 | 3 |
| | EMCI | 5 | 109 | 11 |
| | SMC | 4 | 12 | 93 |

The evaluation results and confusion matrix obtained above show that the accuracy of the random forest model for the training set is 96.9%, which is slightly lower than that of the decision tree model in comparison; the accuracy of the cross-validation set and the test set is slightly higher than that of the decision tree model, with results of 90.8% and 91.2%, respectively. The accuracy of predicting LMCI in the confusion matrix was obtained as 95.7%, 87.2%% for predicting EMCI, and 86.9% for predicting SMC. Therefore, it can be obtained that the classification of the three subclasses (SMC, EMCI and LMCI) included in MCI is good for both random forest and decision tree, and the accuracy of prediction for LMCI is higher than that of SMC and EMCI. for both models the accuracy of prediction for LMCI is better than that of random forest, and the accuracy of prediction for SMC and EMCI is better than that of random forest. Decision tree.

Through the above decision tree classification and random forest classification we can find that the classification results for the three subclasses included in MCI (SMC, EMCI and LMCI) are better than those of the three major classes CN, MCI and AD, and overall they all have good results with test set accuracies above 80%.

## 5.4 Question 4.

### 5.4.1 ARIMA time series model

A time-series approach to the problem is used to study the pattern of importance indicators for different categories of diseases over time. A time series is an

arrangement of the data meeting a certain indicator in the order of their occurrence in time, and the arrangement can be either in the form of annual, quarterly, monthly time or other time forms. These series present themselves by random chance due to random factors, and there are dependencies between the members of the series. The series as a whole exhibit dynamic characteristics between the series, which we call the memorability of the time series, as shown by the fact that the data can show some correlation even if the time interval is large. The advantage of this type of data is that it takes advantage of the intrinsic laws of the data for modeling. The most commonly used model for fitting time series is the summated autoregressive moving average model, i.e., the $ARIMA(p,d,q)$ model. $ARIMA(p,d,q)$ model can model both smooth time series and non-smooth time series after differencing.

$ARIMA(p,d,q)$ The model expressions are.

$$\begin{cases} \Phi(B)\Delta^d Y_t = \theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(X_s \varepsilon_t) = 0, \forall_s < t \end{cases}$$

Where $p$ represents the autoregressive order, $q$ is the moving average order, and $d$ is the difference order performed to transform the non-stationary series into a stationary series. $ARIMA(p,d,q)$ There are three components, namely $AR$、$I$、$MA$ , where $AR$ represents the autoregressive model, $I$ represents the difference order, and $MA$ represents the moving average model. It can be seen that the $ARIMA(p,d,q)$ model is actually a combination of the $AR$ model and the $MA$ model. In particular, when $d = 0$ is used, the $ARIMA(p,d,q)$ model is the $ARIMA(p,d,q)$ model.

## 5.4.2 Analysis and solution of the evolution of AD over time

With the attached data cleaning, we found that elderly people judged as normal age (CN), patients with subjective memory impairment (SMC), patients with early mild cognitive impairment (EMCI), and patients with late mild cognitive impairment (LMCI) would continue to be physically tested every six months and given new judgments. However, we found that with the evolution of some patient categories over time, new judgments will occur when retesting again. In order to discover the pattern of evolution of different categories of diseases over time, we analyzed the correlations from the same sample of specific diagnoses collected at different time points. From the first question we identified the strength of the correlation between the data features and the diagnosis of Alzheimer's disease, so the two features with the strongest correlation were selected to study the analysis of the evolution of the disease categories over time in terms of the change of the different features over time.

Following the operation of time series analysis, we selected the previously highly correlated feature variables LDELTOTAL and mPACCDigit and did time series analysis on the different feature variables to predict.

### 5.4.3 Results Presentation

Based on the variables LDELTOTAL, after autocorrelation, partial autocorrelation, and residual analysis, we selected the model results as ARIMA model (0,1,0), and Figure - represents the original data plot, model fitted values, and model predicted values for this time series model.


Figure 5-6.LDELTOTAL time series chart

and the projected change in LDELTOTAL for the next three years, as shown in Table -

Table 5-13.Time Series Forecast Table

| Predicted value | |
|---|---|
| Order (time) | Predicted results |
| 1 | 14.26086956521739 |
| 2 | 14.521739130434781 |
| 3 | 14.782608695652172 |
| 4 | 15.043478260869563 |
| 5 | 15.304347826086953 |
| 6 | 15.565217391304344 |
| 7 | 15.826086956521735 |

The variables mPACCdigit, after autocorrelation, partial autocorrelation, and residual analysis, the model results for the ARIMA model (0,0,2) test table and based on 1 difference data, Figure - represents the original data plot, model fitted values, and model predicted values for this time series model.


Figure 5-7. mPACCdigit time series plot

and the projected change in mPACCdigit for the next three years, as shown in Table -

Table 5-14. Time Series Forecast Table

| Predicted value | |
|---|---|
| Order (time) | Predicted results |
| 1 | -1.0010869599657934 |
| 2 | -0.8770299189317325 |
| 3 | -0.696644095232552 |
| 4 | -0.696644095232552 |
| 5 | -0.696644095232552 |
| 6 | -0.696644095232552 |
| 7 | -0.696644095232552 |

Combining the data analysis of time series and characteristic variables, people with low LDELTOTAL values will have high cognitive ability and probability of developing Alzheimer's disease, but as LDELTOTAL values increase over time, the elderly with mild cognitive impairment are re-diagnosed as normal elderly. Also, the mPACCdigit value tends to be around 0 over time, which is a sign that the patient is approaching normal.

## 5.5 Question 5

### 5.5.1 The five staging of AD

It is internationally recognized that there are 3 stages of AD: preclinical (CN), early clinical (MCI), and AD dementia. Preclinical and early clinical stages are the best time for treatment.

Stage 1:Preclinical, asymptomatic brain amyloid than ah retrograde, the relevant

biological indicators are positive positron emission computed tomography (PET) or cerebrospinal fluid $\beta-$ amyloid ( $A\beta$ ), positive Tau protein associated with neuronal damage, memory impairment may occur, highlighted by memory loss, prone to fatigue and exertion, anxiety and negative mood, no cognitive function changes for the time being.

Stage 2:Early clinical stage, also known as MCI stage, this stage will be divided into SMC, EMCI and LMCI in this paper, which develops into early neuronal metamorphosis, $A\beta$ and Tau protein positivity, mild changes in cognitive function, increased memory impairment, decreased logical thinking, integrated analysis, repetitive speech, and decreased computational power.

Stage 3:AD dementia with cognitive or behavioral impairment, $A\beta$ and Tau protein positivity, bitter smile wuchang, emotional indifference, loss of language skills, resulting in inability to perform simple daily living tasks such as dressing, walking, eating, etc., and loss of the ability to engage with the outside world.

Depending on the brain areas involved in different stages of the disease, AD patients show different clinical symptoms.

### 5.5.2 Advances in the diagnosis of Alzheimer's disease

In 1984, the National Institute of Neurological Disorders and Stroke - Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) published the first internationally recognized diagnostic criteria for AD, known as the "NINCDS-ADRDA Diagnostic Criteria ".

The diagnostic criteria consist of two parts: the diagnostic criteria for dementia and the diagnostic criteria for AD, which are divided into three levels: "probable AD", "possible AD" and "definite AD". definite AD".

Since there were no biological markers available at the time, the physician's diagnosis relied on medical history, clinical experience and neuropsychological testing. In the diagnostic process, the physician diagnosed dementia mainly through history taking, clinical examination, and neuropsychological assessment, and then excluded other causes of dementia, such as cerebrovascular disease, neurological tumors, and other diseases, before finally diagnosing "probable AD" or "possible AD". possible AD". To make a diagnosis of "definite AD", an autopsy should be performed after the death of the patient and only after finding AD-specific pathological changes such as senile plaques and neuronal tangles.

The NINCDS-ADRDA diagnostic criteria have been proven to have a sensitivity of 81% and specificity of 70% for "possible AD" through numerous clinicopathological studies. The core of the NINCDS-ADRDA criteria for AD diagnosis is the exclusion method, but it still has some drawbacks, such as it only considers AD as a subtype of dementia and does not include the prodromal and

asymptomatic phases of AD, nor does it distinguish AD from other types of dementia. It is easy to misdiagnose non-AD as AD cases, thus affecting the accuracy, rationality and scientific validity of relevant scientific studies.

In 2007, the IWG refined the NINCDS-ADRDA diagnostic criteria, and this new criterion has two significant advances: treating AD as a dynamic process of developmental change and incorporating biological markers into the AD diagnostic criteria for the first time. First, the IWG criteria no longer treat AD simply as an independent disease unit, but as a continuous disease process. For the first time, the concept of MCI-an intermediate state between normal aging and dementia-was introduced, and MCI was classified as an AD diagnosis.

In 2010, on the basis of this version of the diagnostic criteria, the IWG further proposed several emerging concepts regarding AD, for example: preclinical, prodromal AD, atypicalAD, Pathophysiological markers, and topographical markers, etc. The new IWG criteria suggest that preclinical AD includes two states, asymptomatic high-risk state AD and presymptomatic AD, the former refers to the presence of biomarkers of Alzheimer's disease pathology without clinical signs or symptoms, and the latter refers to individuals who eventually carry a single gene mutation in AD that is completely dominant and inevitably manifest AD symptoms clinically.

In 2011, the National Institute on Aging and the Alzheimer's Association (NIA-AA) released a new diagnostic criteria for AD, the "NIA-AA Diagnostic Criteria". The NIA-AA criteria have two distinctive features: first, the preclinical asymptomatic stage of AD is included in AD, which makes the diagnosis of AD much more advanced; second, more emphasis is placed on the value of biological markers for diagnosis, and new biological markers are expanded. The second is that more attention is paid to the value of biological markers for diagnosis, and new biological markers have been expanded. In the NIA-AA diagnostic criteria, the biological markers are divided into two categories. The first category is biological markers reflecting amyloid accumulation, including reduced $A\beta_{42}$ in cerebrospinal fluid and abnormal amyloid trace retention by PET; the second category is biological markers reflecting neuronal damage, including elevated tau protein in cerebrospinal fluid, reduced fluorinated deoxyglucose uptake in temporal pallidum on PET scan, and temporoparietal atrophy on structural MRI scan. Depending on the results of these two types of biological markers, preclinical AD can be classified as "highly likely", "moderately likely", and "unlikely".

In 2014, the IWG once again revised and released new diagnostic criteria, the "IWG-2 Diagnostic Criteria". This criterion still adheres to the diagnostic criteria that AD can be diagnosed by meeting one core clinical diagnostic criterion and at least one pathologically relevant biological marker change in AD. IWG-2 further refined the classification and weighting of biomarkers. $A\beta_{1-42}$ cannot be used as a diagnostic marker alone, but must be combined with T-tau or P-tau, whereas the NIA-AA criteria

only emphasize the different meanings of $A\beta$ The NIA-AA criteria only emphasize the different meanings of tau and tau, but give the same status. It also provides the first division of biological markers into diagnostic and progression markers, with structural MRI and FDG-PET as markers of disease progression that can be used to predict the transition from MCI to AD, and the inclusion of pathogenic gene mutations as diagnostic markers. IWG-2 provides a unified two-dimensional diagnostic approach that allows the use of the same diagnostic criteria without regard to the severity of cognitive impairment. This provides a uniform criterion that applies to all clinical stages throughout the continuum of

In 2019, researchers from the University of Minnesota compared broad three-shot changes in the retinas of Alzheimer's disease patients with healthy individuals, and by scanning the patients' eyes found a small amount of a protein that can cluster and form plaques in the brain, an early sign of disease progression. Further experiments have shown that this technique has a higher sensitivity in the early stages of AD and may provide a huge boost to the treatment of AD.

This paper provides some feasible approaches to the diagnosis of AD based on some methods of machine learning and the use of time series models in statistics using structural brain features and cognitive-behavioral characteristics data, but there is still much room for improvement at present. How sensitive and specific are the current biological markers? Are the new criteria really applicable to clinical practice? With the incidence of AD increasing year by year, these challenges need to be overcome.

Table 5-15

| Category | Classification Method | Processing suggestions |
|---|---|---|
| Absolute risk group | Carriers of autosomal dominant mutations (genes such as amyloid precursor protein, progerin 1 and 2); trisomy 21 | Close monitoring, early treatment or participation in drug clinical trials |
| High Risk Groups | Cerebrospinal fluid or PET detected $A\beta$ protein positive and Tau protein positive; PET borderline extracortical Tau protein positive (Break grading of 5 or higher); $APOE\varepsilon4$ pure haplotypes | Regular follow-up to consider initiating treatment or participating in preclinical drug clinical trials |
| Undetermined risk group | Has an incomplete pattern of biological markers ( $A\beta$ positive, Tau protein negative or unknown; $A\beta$ negative, Tau | Regular follow-up, even if the examination is perfected, to clarify the diagnosis |

protein negative)

Note: $APOE\varepsilon 4$ ,Apolipoprotein E allele $\varepsilon 4$

# 6. Evaluation and optimization of the model

## 6.1 Advantages of the model

This paper uses three models, decision trees, support vector machines, and random forests, and all have a sound theoretical foundation. The decision tree can handle data sets with both quantitative and qualitative variables, and can validate the model using numerical statistical tests, providing a basis for the validation of the explanatory model. Support vector machines have many advantages in solving small samples, nonlinear and high-dimensional data, and they do not need to depend on the whole data, so they have a high generalization ability. Random forests are able to handle high-dimensional data and are not prone to overfitting, and can also determine the importance of features.

## 6.2 Disadvantages of the model

The decision tree model tends to produce an overly complex model, which will greatly reduce the generalization ability of the model, so the minimum number of samples required for the leaf nodes or the maximum depth of the number needs to be set to avoid overfitting. Support vector machines tend to become less efficient when the number of observed samples keeps getting larger, and are more sensitive to missing data. Although random forests are generally less prone to overfitting, they can be overfitted for certain noisy classification and regression problems; and they have a greater impact on random forests when there are more class divisions for attribute variables with different values.

# References


[1] ZHANG Ningyuan,ZHENG Xijun,XU Ling,LIU Hongxia,ZHENG Qingshan. Disease progression model and research progress of Alzheimer's disease[J].Chinese Clinical Pharmacology and Therapeutics,26(06):687-694,2021.

[2] LI Cai,FAN Jiao. Classification prediction of Alzheimer's disease based on machine learning[J].Chinese Journal of Medical Physics,37(03):379-384,2020.

[3] JIANG Peng. The new model predicts the risk of Alzheimer's disease ten years in advance[N].WenWeiPo,2022-09-26(004).DOI:10.28814/n.cnki.nwehu.2022.003069.

[4] YANG Bangkun,WANG Lesheng,NIE Ying,XIONG Wenping. Initial behavior recognition method of Alzheimer's disease based on machine learning[J].Biomedical engineering research,2021,40(02):121-125.DOI:10.19529/j.cnki.1672-6278.2021.02.03.

[5] Sun Mengsha, Gu Mingmin. Research progress in early diagnosis of Alzheimer's disease [J]. Chinese Journal of Modern Neurology,18(03):213-221,2018.

[6] Ding Saineng, Ma Xiaoxi, Zhao Qianhua. Changes in the diagnostic criteria of Alzheimer's disease: interpretation of the 2021 edition of the International Working Group standard [J]. Journal of Neurology and Neurorehabilitation,2021,17(04):135-139.

[7] Mei Yu, Sun Xuwen, Ba Maowen, Yu Guoping. New advances in clinical symptoms and diagnostic methods for different stages of Alzheimer's disease[J]. Modern medicine and health,2017,33(18):2745-2748+2754.

[8] Lin, Luo. The past life and present life of diagnostic criteria for Alzheimer's disease[J]. Science News,2020(05):44-47.

[9] Edward Challis,Peter Hurley,Laura Serra,Marco Bozzali,Seb Oliver,Mara Cercignani.Gaussian process classification of Alzheimer’s disease and mild cognitive impairment from resting-state fMRI [J]. NeuroImage . 2015

[10] Babak A. Ardekani,Elaine Bermudez,Asim M. Mubeen,Alvin H . Bachman.Prediction of Incipient Alzheimer’s Disease Dementia in Patients with Mild Cognitive Impairment [J]. Journal of Alzheimer’s Disease . 2016 (1)

[11] Zeng An,Jia Longfei ,Pan Dan,Song Xiaowei.Early diagnosis of Alzheimer's disease based on convolutional neural network and integrated learning[J]. Journal of Biomedical Engineering,2019,36(05):711-719.

[12] Yang Zhiyong, Wang Huaqiao, Yao Zhibin. Quantitative analysis of psycho-behavioral symptoms in patients with Alzheimer's disease[J]. Chinese Journal of Gerontology,2011,31(03):408-410.

[13] Peng Guoping, Liu Xiaoyan, Luo Benyan. The role of neuroimaging in the diagnosis and differentiation of Alzheimer's disease[J]. Chinese Journal of Neurology,2020(05):321-322-323-324-325-326-327.

# Appendix

```
from sklearn.ensemble import RandomForestClassifier as RandomForest
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import scipy as sp
import seaborn.apionly as sns
import sklearn.metrics as metrics
import warnings
adnimerge = pd.read_csv('E:/program-py/sss.csv')
adnimerge.head()
adnimerge_bl = adnimerge.loc[adnimerge['VISCODE']=='bl']
adnimerge_bl = adnimerge_bl.drop('M', axis=1)
adnimerge_bl    =    adnimerge_bl.drop(['EXAMDATE_bl',    'CDRSB_bl',
'ADAS11_bl', 'ADAS13_bl', 'MMSE_bl',
      'RAVLT_immediate_bl',                    'RAVLT_learning_bl',
'RAVLT_forgetting_bl',
      'RAVLT_perc_forgetting_bl', 'FAQ_bl', 'Ventricles_bl',
      'Hippocampus_bl',      'WholeBrain_bl',      'Entorhinal_bl',
'Fusiform_bl',
      'MidTemp_bl',     'ICV_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',
'FDG_bl',
      'AV45_bl', 'Years_bl', 'Month_bl'], axis=1)
adnimerge_bl.head()
adnimerge_bl['DX_bl'].value_counts()
df1=adnimerge_bl['PTRACCAT'].value_counts()
df1
adnimerge_bl.shape
for col in adnimerge_bl.columns:
    print(col, adnimerge_bl[col].isnull().sum())
data1=adnimerge_bl[[
'MMSE','RAVLT_immediate','RAVLT_learning','RAVLT_forgetting',
'RAVLT_perc_forgetting','LDELTOTAL','FAQ','Ventricles','DX_bl''WholeB
rain','ICV','']]
data1
```

```
for col in data1.columns:
    print(col, data1[col].isnull().sum())
cols_with_na=[]
for column in data1:
    if data1[column].isnull().any()==True:
        cols_with_na.append(column)
meandlf = data1.copy()
for col in cols_with_na:
    if data1[col].dtype=='float64':
        meandf[col] = data1[col].fillna(data1[col].mean())
    else:
        meandf[col] = data1[col].fillna(data1[col].mode()[0])
np.any(pd.isnull(meandf))
meandf.head()
meandf.isnull().sum()
mapping = {
           'CN': 0,
           'AD': 1,
           'LMCI': 1,
           'EMCI':1,
           'SMC':1
}
meandf['DX_bl'] =meandf['DX_bl'].map(mapping)
meandf.head()
meandf.isnull().sum()
X=meandf.drop('DX_bl',axis=1)
y=meandf['DX_bl']
X
plt.bar(y.value_counts().index, y.value_counts())
plt.show()
X_train,    X_test,    y_train,    y_test    =    train_test_split(X,
y,test_size=0.2,random_state=0)
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=50)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
from sklearn.metrics import classification_report
result1 = classification_report(y_test, y_pred)
print("Classification Report:",)
print (result1)
```

```
result2 = accuracy_score(y_test,y_pred)
print("Accuracy:",result2)
importances = classifier.feature_importances_
feat_labels =  meandf.columns[:-1]
indices = np.argsort(importances)[::-1]
for f in range(X_train.shape[1]):
    print("%2d) %-*s %f" % (f + 1, 30, feat_labels[indices[f]],
importances[indices[f]]))
importances = classifier.feature_importances_
print(importances)
x_columns =meandf.columns[1:]
indices = np.argsort(importances)[::-1]
x_columns_indices = []
for f in range(X_train.shape[1]):
    print("%2d) %-*s %f" % (f + 1, 30, feat_labels[indices[f]],
importances[indices[f]]))
    x_columns_indices.append(feat_labels[indices[f]])
print(x_columns_indices)
print(x_columns.shape[0])
print(x_columns)
print(np.arange(x_columns.shape[0]))
# ## XGBOOST
import xgboost as xgb
from sklearn.model_selection import RandomizedSearchCV
# xgboost
from xgboost import XGBClassifier
model = XGBClassifier()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
y_pred
res = classification_report(y_test, y_pred)
print("Classification Report:",)
print (res)
res2 = accuracy_score(y_test,y_pred)
print("Accuracy:",res2)
from sklearn.model_selection import KFold
param_dist = {
        'n_estimators':range(80,200,4),
        'max_depth':range(2,15,1),
        'learning_rate':np.linspace(0.01,2,20),
        'subsample':np.linspace(0.7,0.9,20),
        'colsample_bytree':np.linspace(0.5,0.98,10),
        'min_child_weight':range(1,9,1)
        }
```

```
cvv = KFold(3)
grid = RandomizedSearchCV(model,param_dist,cv = cvv,n_iter=100,refit="roc_auc",n_jobs = -1)
grid.fit(X_train,y_train)
best_estimator = grid.best_estimator_
print(best_estimator)
print(grid.best_score_)
# SVM
from sklearn.svm import SVC
model = SVC()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
print(classification_report(y_test, predictions))
from sklearn.model_selection import GridSearchCV
param_grid = {'C': [0.1, 1, 10, 100],
              'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
              'gamma':['scale', 'auto'],
              'kernel': ['linear']}
grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3,n_jobs=-1)
grid.fit(X_train, y_train)
print(grid.best_params_)
grid_predictions = grid.predict(X_test)
 #ARIMA
import numpy
import pandas
from spsspro.algorithm import statistical_model_analysis
data = pandas.DataFrame({
    "A": numpy.random.random(size=20)
})
result = statistical_model_analysis.arima_analysis(data=data, p=0, d=0, q=0, forecast_num=10)
print(result)
```