Team Number :	2755

Problem Chosen :	A

2017 APMCM summary sheet

Effects of Sleep on Human Body

Summary

This paper studies the effects of sleep on human body, taking into account the correlation between sleep and various indicators, giving valid and reasonable suggestions.

For the first requirement, analysis of Annex I and sleep quality data, as there may be some correlation between the six indicators, so the use of first-order partial differential method, after dealing with these data, finding that the correlation coefficient between the two indicators are less than 0.85, it can be regarded as the correlation between indicators is weak, in line with the gradual regression data analysis. Because not every index is related to the Sleep Quality, Sleep Quality is correlated with Age, Sex, Nervousness and Charater by stepwise regression analysis, it isn't correlated with Reliability and Psychoticism.

For the second requirement, analysis of the relationship between sleep and diagnosis results, there are 278 kinds of statistical symptoms, and the parameters of sleep-related indicators are all four discrete values, which are 0, 1, 2, and 3, each disease is also a discrete and individual. Therefore, the problem can be classified into categories problem. By establishing a decision tree model and using Python software to calculate, a depth of 3 dendrogram between sleep and illness was obtained, it is the relationship between the diagnosis results and sleep.

For the third requirement, ten patients were diagnosed with a decision tree model based on Problem Two. 80% of patients were diagnosed with Sleep Disorder, and 20% of patients were diagnosed with Depression. Both of these conditions are also the highest among the 278 conditions mentioned in the sample.

For the fourth requirement, this paper combines with the results of the second and third studies, introduced of Pittsburgh Sleep Quality Index(PSQI), PSQI score greater than or equal to 8 points as a plan for good and poor quality of sleep, respectively, poor quality of sleep, the current good quality of sleep were put forward reasonable sleep recommendations. Evaluation of sleep planning use Analytic Hierarchy Process(AHP) and acquired seven factors weight coefficient, the two kinds of sleep plans to enhance the efficiency is 75.1%, 53.1% respectively ,we can find that the improvement of the sleep program is more obvious.

Finally, this paper introduce decision tree model to deal with a large number of sample indicators, so that the results obtained by processing the data are more accurate and reliable.

Key words: stepwise regression decision tree Python dendrogram PSQI AHP

Contents

1. Restatement of the problem

1.1 Background introduction

Sleep is an universal phenomenon necessary for maintaining homeostasis and function across a range of organs, lack of sleep has severe health-related consequences affecting whole-body functioning [1]. Since 2001, the World Association of Sleep Medicine has set March 21 every year as world sleep day for drawing attention to the important and quality of sleep. The mental state of a whole day depends on the sleep quality last night, what's more, high sleep quality naturally ensures people to be energetic. According to statistics, however, the rate of insomnia in Chinese adults is as high as 38.2%, and the rate of insomnia in adolescents in rising meanwhile. In general, it defined as insomnia if the time to fall asleep is more than 30 minutes, thus we think many participants are also the insomniac. Insomnia symptoms are prevalent in the general population and associated with a variety of negative outcomes [2]. long-term insomnia makes people feel tired. Lack of energy in the whole day, and cannot focus attentions, so the efficiency of working and studying are obviously low. Severe insomnia even will cause autonomic nerve function disorder, leading to imbalance and various systems in the body.

Many factors bring about insomnia, including objective and subjective factors. Objective factors are environmental changes, a cup of tea or coffee before going to bed, and so on, and subjective factors are generally the pressure of life, emotional loss, mental excitement and other spiritual factors. However, the brains of young people in the period of growth and development are extremely prone to fatigue due to learning and work stress. Therefore, they have to pay special attention to bed rest to ensure a good and healthy body. The study of the sleep tend to a profound influence on the surrounding, either in China or throughout the world have important reference and guiding significance.

1.2 The problem to be solved

(1) Analyze the relationship between the indicators given and the quality of sleep according to the date in Annex I (Source and test number are not indicators), if there is no correlation between one or several indicators and sleep quality, find it or them out and delete.

(2) Analyze the relationship between the diagnosis results and sleep, The relevant scores for sleep conditions are given in Annex II ("0" for good, "1" for normal, "2" for poor, and "3" for very poor), the higher the score, the worse the sleep condition.

(3) Assuming you are a doctor, what diagnosis would you make to the patient based on the date in Annex III? Give your diagnosis result.

(4) How to scientifically arrange our rest time for the health of the body? Develop appropriate sleep program and evaluate its effectiveness.

2. Problem Analysis

Studying the effect of sleep on human body in daily life has a very important guiding significance. The purpose of this research is to study the indicators that affect the quality of sleep and the relationship between poor sleep habits and diseases, and propose positive and effective suggestions to achieve favorable sleep.

2.1 Analysis of problem 1

Based on 6349 sets of data, the correlation between the indicators given in Annex I and sleep quality was analyzed. From the dry, we can see, Source and test number is not a indicators affecting the quality of sleep, which means that we only need to analyze the relationship between the other six factors and sleep quality. Because six indicators do not necessarily have an impact on the quality of sleep, and there may be a strong correlation between the various indicators. Therefore, the first-order partial correlation analysis was used to analyze the correlation between the six factors and verify the independence between the variables of each factor. Due to the stepwise regression analysis, it is suitable to use under the condition that the correlation between each variable is not strong. If the independence is strong, stepwise regression analysis is used to select the indicators that have a significant impact on sleep quality from the six variables. If the independence is not strong, consider other means of data analysis.

2.2 Analysis of problem 2

Based on the data in Annex II, we analyzed the relationship between sleep-related seven indicators and diagnosis results. The data were classified and co-ordinated, of which there were 278 kinds of diseases, and the disease was not a series of continuous values. The relevant data of seven indicators were discrete independent values, but is simply "0", "1", "2", "3". Therefore, the problem can be considered as a classification problem, through the development of a series of rules to the data classification.

2.3 Analysis of problem 3

Annex III gives 10 patient data, based on the data given, diagnoses each patient's corresponding condition. Since the relevant indicators given in Annex III are the same as the variables in Model II, it is possible to diagnose the condition of each patient based on the model of Problem 2.

2.4 Analysis of problem 4

This question is based on the results of the previous analysis, put forward suggestions related to sleep and evaluate the rationality of the proposal. A statistical analysis of the sleep-related parameters in Annex II. Generally assessed the condition of sleep in the sample. 6438 samples are divided into two types by the method of Pittsburgh Sleep Quality Index (PSQI), seven indicators in Annex II and the PSQI score greater than 8 points, the subjective sleep quality is poor, on the contrary, the subjective sleep quality is good. For different conditions of sleep, making different recommendations.

3. Assumptions

(1) Assuming that the majority of the data given in the Annex is true and reliable, the possibility of a few abnormal data is not ruled out;

(2) When the correlation coefficient between two indicators is small, they are considered as independent of each other;

(3) Not consider the small probability of illness;

(4) Ignoring the impact of culling data on the analysis results;

(5) Assuming there is no interaction between the various diagnostic results;

(6) Not consider the diagnosis that does not appear in the schedule of the results.

4. Symbol explanations

Symbol	Explanation
x_1	Age
x_2	Sex
x_3	Reliability
x_4	Psychoticism
x_5	Nervousness
x_6	Character
Y	Sleep quality
H(p)	The entropy of Q
D	Training data sets
$\|D\|$	Sample size
$H_t(T)$	Empirical entropy on leaf node t
g(D,A)	Feature A Information gain on training data set D.
D_{ik}	Set of samples belonging to class M_K in subset D_i
$\|T\|$	Tree T leaves the number of nodes
$\|M_K\|$	The number of samples belonging to class M_K
$\|D_{ik}\|$	the number of samples D_{ik}
H(Z|Q)	the uncertainty of random variable Z under the condition of Q of random variables
w_1	Subjective sleep quality
w_2	Sleep latency
w_3	Sleep time
w_4	Sleep efficiency
w_5	Hypnagogue
w_6	Sleep disorder
w_7	Daytime dyfunction

5. The solution of the problem

5.1 The solution of problem 1

(1) First-order partial correlation analysis

Through the analysis of the problem, we know that there are 6 independent variables: Age, Sex, Reliability, Psychoticism, Nervousness and Character, they are defined as $x_1, x_2, x_3, x_4, x_5, x_6$. First, we determine the independence of the independent variables and use first-order partial correlation analysis the above variables by SPSS software. Below is the results in Table 1.

Table 1 six indicators relatedness

Control variable			Age	Sex	Reli	Psych	Nerv	Char
Sleep quality	Age	correlation	1.000	-.041	.385	.203	-.264	.085
		signification		.001	.000	.000	.000	.000
		df	0	6346	6346	6346	6346	6346
	Sex	correlation	-.041	1.000	-.047	-.349	.048	-.059
		signification	.001	.	.000	.000	.000	.000
		df	6346	0	6346	6346	6346	6346
	Reli	correlation	.385	-.047	1.000	-.184	-.358	-.001
		signification	.000	.000	.	.000	.000	.909
		df	6346	6346	0	6346	6346	6346
	Psych	correlation	.203	-.349	-.184	1.000	.032	.021
		signification	.000	.000	.000	.	.010	.092
		df	6346	6346	6346	0	6346	6346
	Nerv	correlation	-.264	.048	-.358	.032	1.000	-.183
		signification	.000	.000	.000	.010	.	.000
		df	6346	6346	6346	6346	0	6346
	Char	correlation	.085	-.059	-.001	.021	-.183	1.000
		signification	.000	.000	.909	.092	.000	.
		df	6346	6346	6346	6346	6346	0

Note：Reliability--Reli, Psychoticism--Phych, Nervousness--Nerv, Character--Char.

The first-order partial correlation analysis is to calculate the correlation between variables under the influence of the absence of other relevant indicators. Here to analysis the correlation between variables under controling the linear effect of sleep quality conditions.

Table 1 shows that the correlation coefficient between each two variables. Generally thinking, there is a significant correlation of each two variables when the correlation coefficient [3] between them is greater than 0.85. but there are six variables, and correlation coefficient of every two of them is less 0.85. So the correlation between the variables is weak and can’t affect each other, so the problem meets the requirements of stepwise regression analysis.

(2) Stepwise regression analysis

The basic idea of step by step analysis: Firstly, determine an initial set containing several independent variables, secondly, introduce one variable that has the most effect on the variable to above set from one outside the set, third test the new set, finally remove one variable that has the least effect from the variable that becomes insignificant. Do this until it can't be introduced and removed. Both the introduction and the removal are based on a given level of significance.

The data in Annex I will be processed and some data will be intercepted in Table 2. Step-by-step analysis of the data using the stepwise (x, y) that is Matlab function and get the initial window of stepwise Regression, shown in Figure 1

Table 2 partially intercepted data after processing

Factor / Test number	Age	Sex	Reliability	Psychoticism	Nervousness	Character	Sleep quality
1	28	1	36.62	15	63.91	50.34	1
2	65	0	57.24	44.2	61.17	37.7	2
3	63	0	48.45	40.99	41.73	35.35	1
4	30	0	52.42	40.48	67.9	28.41	3
5	67	0	54.31	37.79	52.53	42.39	1
6	53	1	36.05	62.95	66.81	51.02	2
7	26	1	64.79	4	75.15	48.05	2
8	61	1	33.55	54.59	63.85	28.18	3
9	19	0	36.85	44.05	77.89	48.48	3
10	56	0	47.89	54.05	59.29	64.56	3
11	75	1	36.89	62.11	77.36	37.41	1
12	71	0	30.85	63.43	65.49	54.13	3
13	34	1	54.94	45.68	26.59	50.12	0
14	62	1	50.27	58.35	79.62	55.89	3
15	20	0	36.85	49.88	77.89	32.39	3
16	45	0	60.75	47.47	31.75	55.13	2
17	43	0	57.75	47.47	56.81	40.81	2
18	27	1	33.8	9	68.4	61.81	1
19	25	0	43.57	53.22	74.54	30.82	3
20	50	0	53.24	45.12	38.41	32.54	2
…	…	…	…	…	…	…	…

Note: the male is "1", the female is "0"

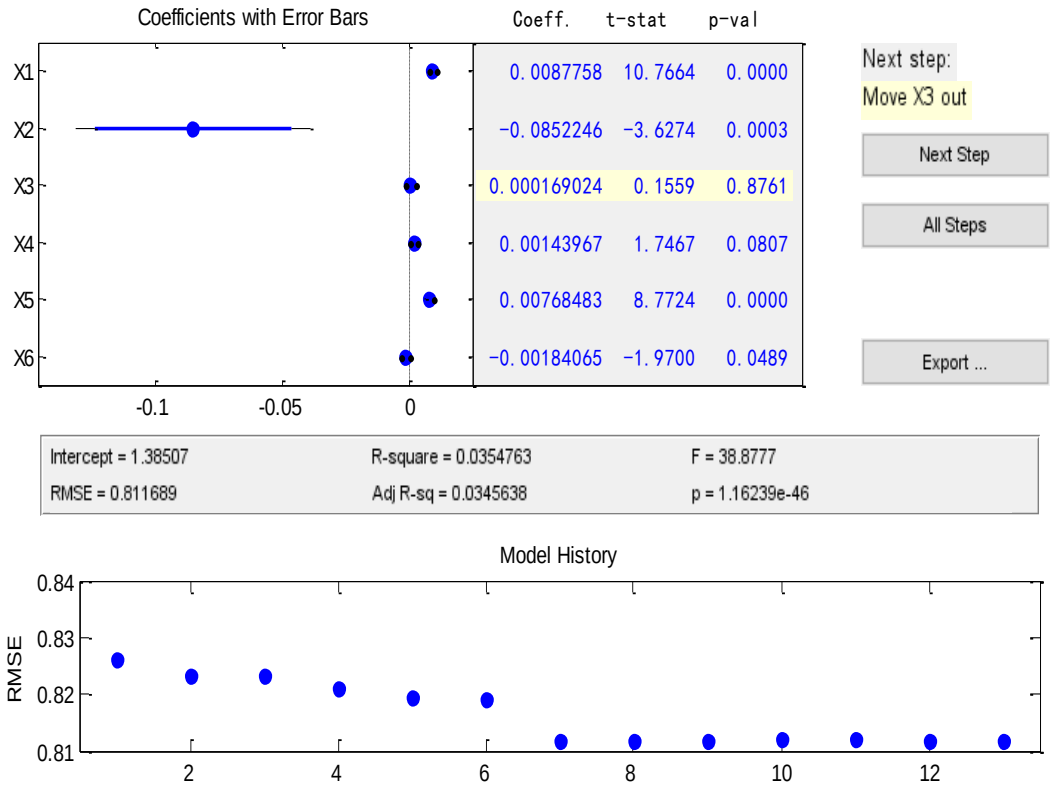

Figure 1 Stepwise Regression initial interface

The upper left corner of the interface gives an estimate of the regression coefficients and the error bounds (in horizontal lines) for all six variables, where the colored horizontal line on the computer screen represents the confidence interval of 90%, the gray level is Confidence interval of 95%. If the horizontal line is red, it means that the variable of the horizontal line has not been selected to the model. The table in the middle of the upper part of the screen shows the estimated regression coefficients for the variable, the statistic for the test t, and the p-value once a variable has been selected for the model. In general, what is selected at each step is the item with the smallest p-value or the largest statistic. In Figure 1 is the independent variable x1, generally just press the Next Step button for the next step, the program will automatically select the arguments to be introduced or removed, and the corresponding results are given in the upper right of the interface. Of course, you can also do this manually, by clicking on a row in the table with the mouse to change its state, that is, one of the variables that is not currently in the model (the red row) was introduced (turns blue), currently a variable (blue line) in the model is removed (turns red) only until the screen prompts Move no terms. Usually we can press the All Step button to complete the step-by-step regression of the entire model. Press the All Step button shown in Figure 1 to get the final result of a step-by-step regression.

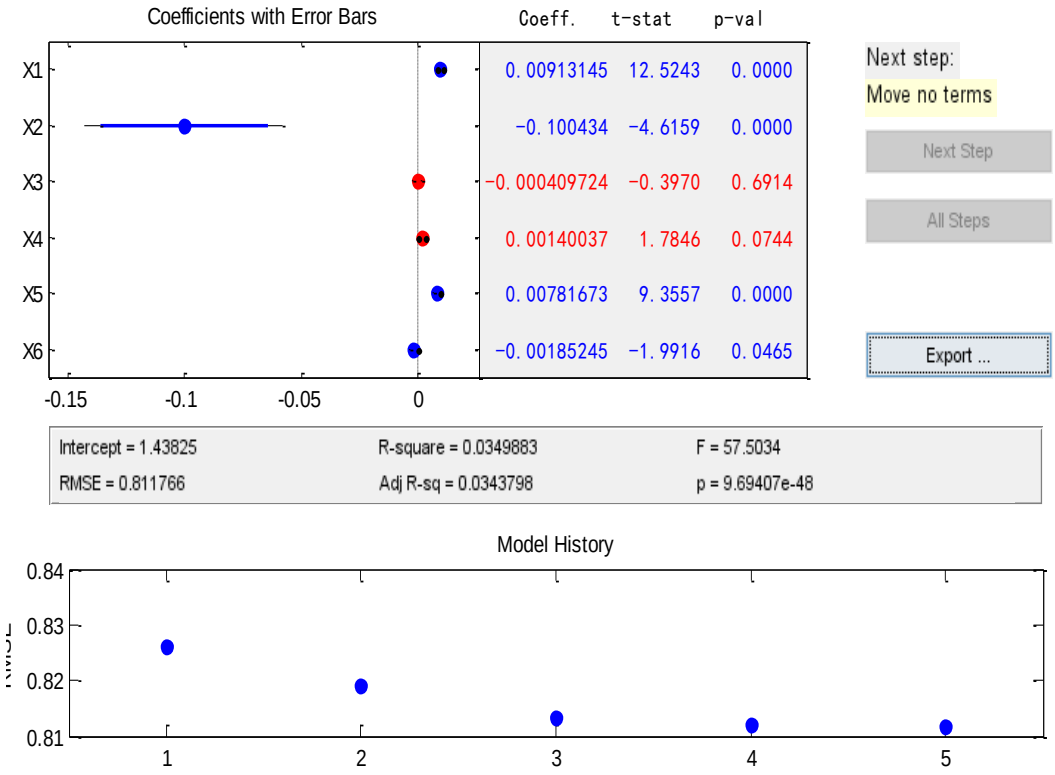

Figure 2 Stepwise Regression final screen (press All Step button income)

Stepwise Regression interface is divided into upper and lower three parts, the upper part of the interface has been introduced, the interface of the middle part of the table gives all the results of the regression model: Intercept, decision coefficient R^2, F, RMSE, adjusted coefficient of determination R^2, and p of the test. And $R^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$, Where k is the number of arguments selected for the model.

The lower half of the interface, Model History, gives the point of the standard deviation of the residuals for each step in the step-by-step regression. Move the mouse to the blue point corresponding to the step and the corresponding part of the model will be displayed Of the independent variable, click on the blue point to track the current model of the corresponding interface.

For our problem, we can get from Figure 2 that the final selected variables are x_1, x_2, x_5, x_6 shown in blue. From the analysis we can know that the residual standard deviation corresponding to each step is basically the same with the gradual regression,. Although the R^2 slightly decreases in individual steps, the F value of the model doubles, which shows that the elimination off x_3, x_4 model is appropriate. The regression coefficients and regression constants with x_1, x_2, x_5, x_6 are obtained by stepwise regression. The final model is:

$$Y = 0.0091x_1 - 0.1004x_2 + 0.0078x_5 - 0.0018x_6 + 1.43825$$

In the final model, the regression variables are x_1, x_2, x_5, x_6 ,it is a simple and easy model, which means that Age, Sex and Nervousness are obtained. Character and Sleep Quality are strongly correlated.

(3) Conclusion

Through stepwise regression analysis, we found that Sleep Quality is related to Age, Sex, Nervousness and Character, but not related to Reliability and Psychoticism. The influence of each index on sleep quality can be expressed by stepwise regression to a certain extent. The influence coefficients of the four indexes on sleep quality are respectively 0.0091, -0.1004, 0.0078 and -0.0018. Said that its impact on the quality of sleep is a positive effect, reverse is negative. The greater the absolute value of the value, the greater its impact, on the contrary, the opposite.

5.2 Model Establishment and Solution of problem 2

Problem 2 can be attributed to the classification problem, and decision tree [4] is a basic classification and regression method, which is a set of if-then rules. This question adopts the method of decision tree to formulate a series of rules and establish the model.

(1) Index selection and data preprocessing

Analysis of the relationship between the diagnosis result and sleep, sleep-related indicators in Annex II are Sleep Quality, Sleep Latency, Sleep Time, Sleep Efficiency, Sleep Disorder, Hypnagogue, Daytime Dyfunction, respectively, so select the seven indicators; out of more than 6,000 samples for pretreatment [5]. Due to incomplete or missing information in some samples, so the number of samples after initial screening may be 6274. After sorting the samples for primary screening, the number of various types of diagnosis results is shown in Table 1 of Appendix. Through the analysis of the data in Annex I, it is found that there are some types of diagnosis results in the diagnosis results, the sample number may be contingent. In order to reduce the chance of the data and improve the accuracy of the data, so corresponding to the number of samples less than 10 to remove the diagnosis, the final number of samples are 5804 after the second screening. Specific as shown in Table 3 below:

Table 3 secondary screening data summary table

Diagnosis	Serial Number	Total Number
Sleep disorder	1	1699
Depression	2	1472
Anxiety disorder	3	862
Anxiety	4	402
Mixed Anxiety And Depression	5	368
Bipolar Affective Disorder	6	132
Non-Organic Insomnia	7	106
Recurrent Depressive Disorder	8	74
Adjustment Disorder	9	58
Anxiety,Sleep disorder	10	50
Schizophrenia	11	47
Sleep disorder,Depression	12	33
Sleep disorder,Mixed Anxiety And Depression	13	33
Mixed Anxiety And Depression Disorder	14	32

Non-Organic Insomnia,Anxiety disorder	15	31
Depression,Sleep disorder	16	28
Sleep disorder,Anxiety	17	28
Somatoform Disorders	18	28
Emotional problem	19	28
Obsessive-Compulsive Disorder	20	28
Sleep disorder,Anxiety disorder	21	26
Mixed Anxiety And Depression,Sleep disorder	22	26
Dysthymia	23	26
Bipolar Affective Disorder,Moderate Eepressive Episode	24	26
Mood Disorder	25	21
Anxiety,Depression	26	20
Somatization Disorder	27	20
Postpartum Depression	28	19
Anxiety disorder,Sleep disorder	29	17
Generalized Anxiety Disorder	30	17
Mental Disorders	31	16
Panic Attacks	32	11
Sleep- Wake Rhythm Disorders	33	10
Anxiety depression	34	10

(2) The establishment of decision tree model

1) Information gain

In information theory and probability statistics, entropy is a measure of the uncertainty of random variables. Let X be a finite number of discrete random variables, the probability distribution is:

$$P(Q = q_i) = p_i \quad i = 1,2, \dots n$$

Then the entropy of random variable X is defined as:

$$H(Q) = -\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

In equation (1), $0\log 0 = 0$ is defined ,if $p_i = 0$. Usually, the logarithm of the base 2 in equation (1) or the base of e (natural logarithm), then the units of entropy are called bits or NAT, respectively. By definition, entropy only depends on the distribution of X, and has nothing to do with the value of X, so the entropy of Q can also be written as H (p)

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

The larger the entropy, the greater the uncertainty of the random variable. Verifying from definition:

$$0 \le H(p) \le \log n$$

When the random variable takes only two values, such as 1,0, the distribution of Q is

$$P(Q = 1) = p \quad P(Q = 0) = 1 - p \quad 0 \le p \le 1$$

Entropy:

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p)$$

When p = 0 or p = 1, H (p) = 0, there is no uncertainty about random variables. When p = 0.5, H (p) = 1, the entropy value is the largest, and the uncertainty of random variables is the largest.

With random variables (Q, Z), the joint probability distribution is

$$P(Q = q_i, Z = z_j) = p_{ij} \quad i = 1,2 \dots, n \quad j = 1,2 \dots, n$$

The conditional entropy H (Z | Q) indicates the uncertainty of the random variable Z under the condition that the random variable Q is known. The conditional entropy H (Z | Q) of the random variable Z given by the random variable Q is defined as the entropy of Q for the conditional probability distribution of Z under given conditions:

$$H(Z|Q) = \sum_{i-1}^{n} p_i H(Z|Q = q_i) \quad (2)$$

When the probability of entropy and conditional entropy is estimated from the data, the corresponding entropy and conditional entropy are respectively called the empirical entropy and the empirical conditional entropy; The information gain indicates that the degree to which the information of the characteristic Q and the uncertainty of the information of the class Z is reduced.

Information gain g (D, A) of training data set D of feature A is defined as the difference between the empirical entropy H (D) of set D and the empirical conditional entropy H (D | A) of D given by feature A. That is, Can be expressed as:

$$g(D, A) = H(D) - H(D|A)$$

The difference between entropy H (Y) and conditional entropy H (Z | Q) is called mutual information. Information gain in decision tree learning is equivalent to household information of classes and characteristics in training data set. Decision Tree learning application information gain criteria selection characteristics. Given the training set dataset D and signature A. empirical entropy H (D) represents the uncertainty of classifying data set D, and empirical conditional entropy H (D | A) represents the uncertainty of classifying data set D under the conditions given by feature A. Then their difference, ie, information gain, indicates the degree to which the uncertainty of the classification of data set D is reduced due to feature A. Obviously, for dataset D, the information gain depends on the feature, different features often have different information gain, and the information gain feature has stronger classification ability.

According to the information gain criterion, the feature selection method is: for the training data set D, calculating the information gain of each feature, and comparing their size, selecting the feature with the largest information gain.

Suppose the training dataset is D, $|D|$represents its sample size, it is the number of samples, with K classes$|M_K|, k = 1,2 \dots, n$, $|M_K|$ is the number of samples belonging to class M_K, $\sum_{k=1}^{K}|M_K| = |D|$, Suppose feature A has n different values $\{a_1, a_2, \dots, a_n\}$, according to the

value of feature A, D is divided into n subsets $D_1, D_2, \ldots, D_n$, $|D_i|$is the number of samples of D_i, $\sum_{i=1}^{n}|D_i| = |D|$, the set of samples belonging to class M_K in subset D_i is D_{ik}, that is $D_{ik} = D_i \cap M_K$, and $|D_{ik}|$ is the number of samples of D_{ik}.

1) The empirical entropy H(D) of computational data set D

$$H(D) = -\sum_{k=1}^{K}\frac{|M_k|}{|D|}\log_2\frac{|M_k|}{|D|}$$

the empirical conditional entropy of characteristic A for data set D is calculated $H(D|A)$

$$H(D|A) = \sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i) = -\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}$$

The computational information gain

$$g(D,A) = H(D) - H(D|A)$$

2) Pruning Decision Tree

Pruning of decision trees is often done by minimizing the tree's loss function or cost function of the decision tree. Suppose the number of leaf nodes in tree T is $|T|$, t is the leaf node of tree T, there are N_t sample points in this leaf node, and there are N_{tk} sample points in k class. k = 1,2, ... K, $H_t(T)$ is the empirical entropy on the leaf node t, $\alpha \geq 0$ is the parameter.

$$C(T) = \sum_{t=1}^{|T|}N_tH_t(T) = -\sum_{t=1}^{|T|}\sum_{k=1}^{K}N_{tk}\log\frac{N_{tk}}{N_t}$$

$$C_\alpha(T) = C(T) + \alpha(T)$$

When α is determined, the larger the subtree is, the better the fitting is with the training data, but the higher the complexity of the model, on the contrary, the smaller sub-tree is, the less fitting the training data. It can be seen that the pruning of the decision tree also takes into account the reduction of the complexity of the model and the learning of the overall model by optimizing the loss function.

3) The solution Model 2

The secondary screening samples were imported into Python software, 80% of the samples were selected as the training set data and the remaining 20% as the prediction set data. The relevant functions were constructed and the decision tree was generated, the decision tree was generated with 14 trees and predicted, and got the accuracy of the data. At the same time, we write an iterative function and obtain the entropy of experience, empirical conditional entropy, entropy of information gain, entropy of information gain to get the influence degree of sleep-related indexes on diagnosis results. As shown in Figure 3 below:

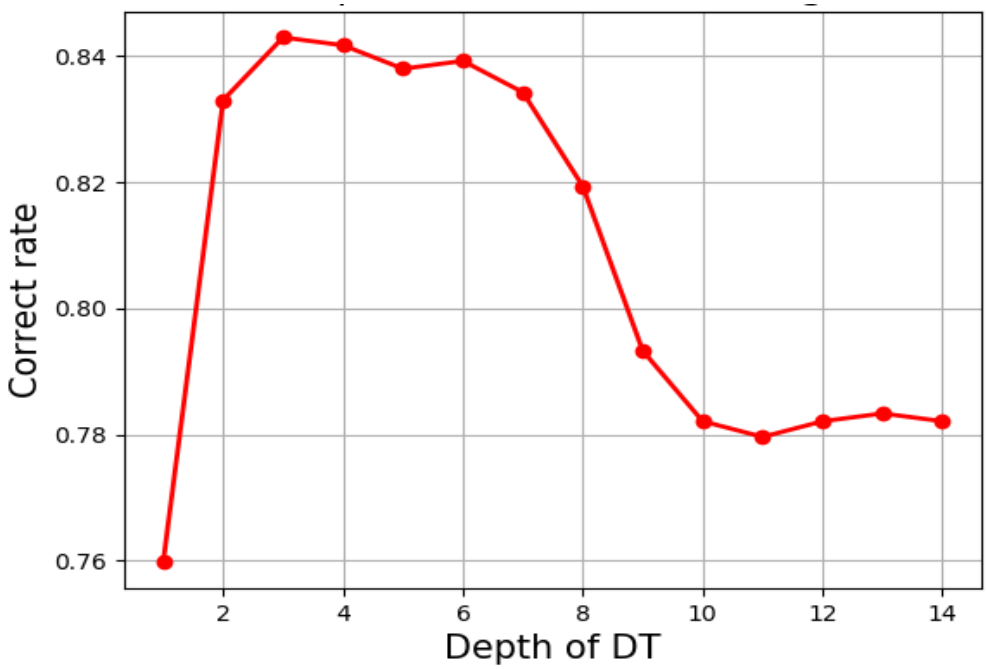

Figure 3 Depth of DT & Over Fitting

The above figure shows that the relationship between the depth of the decision tree and the correctness rate increases first and then declines with the increase of the depth of the decision tree. When the decision number depth is 3, the correct rate reaches the peak value. Therefore, we select a tree with the highest correctness as the overall correctness rate. The specific tree diagram is shown in Figure 4 (see Figure 1 for complete clarity)

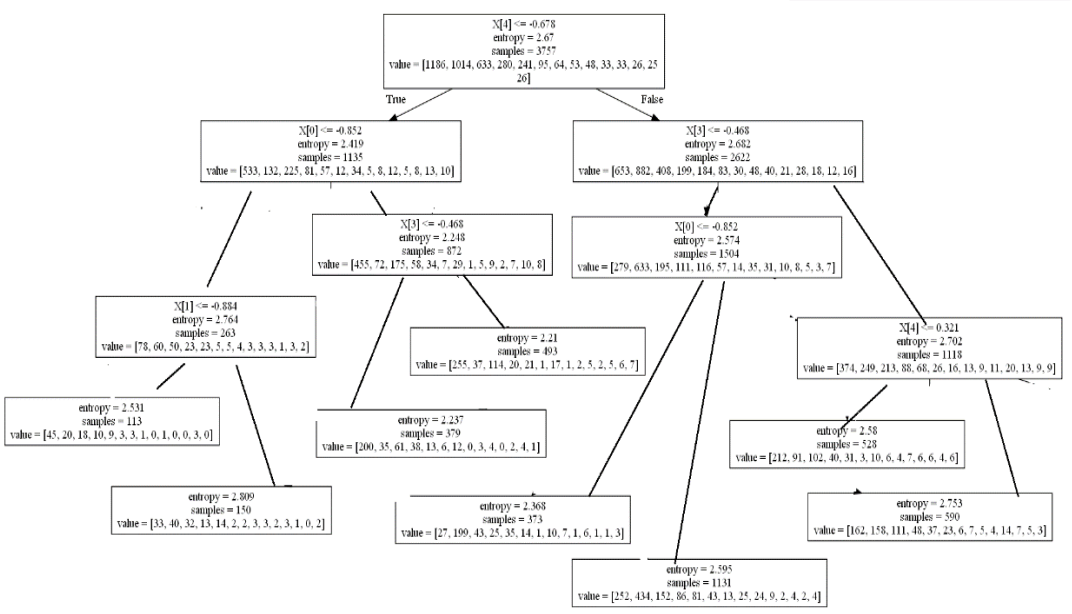

Figure 4 Depth of 3 decision tree dendrogram

Figure 5 shows the specific enlargement

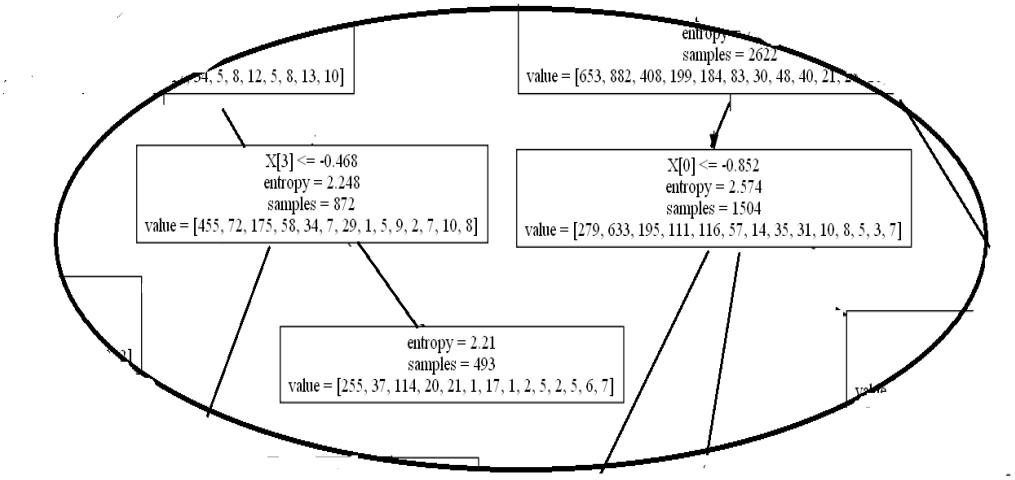

Figure 5 depth of 3 tree part of the decision tree zoom

From Figure 4, the relationship between diagnosis results and sleep. The resulting tree of decision trees represents a series of guidelines that determine the patient's final condition based on the specific characteristics of the patient

5.3 The solution of Problem 3

The patient's condition is diagnosed on the basis of the patient's eigenvalue given in the Annex III. Based on the model and procedure established in Question 2, the rules established by the decision tree are used to diagnose the patient's condition. As shown in Table 4 below.

Table 4 Patient diagnosis results

Test Number	Sleep quality	Sleep latency	Sleep time	Sleep efficiency	Sleep disorder	Hypna-gogue	Daytime dyfunction	diagnosis result
1	1	0	1	0	1	1	1	Sleep disorder
2	2	1	2	1	2	1	0	Sleep disorder
3	3	2	3	2	3	0	0	Sleep disorder
4	3	3	3	0	2	2	2	Depression
5	2	2	2	3	0	3	0	Sleep disorder
6	3	2	2	0	1	3	1	Sleep disorder
7	1	1	1	1	0	2	2	Depression
8	1	3	3	2	1	1	1	Sleep disorder
9	3	2	2	3	3	2	0	Sleep disorder
10	1	2	3	2	1	2	1	Sleep disorder

As shown in Table 4, Sleep disorder accounted for 80% of the 10 patients and 20% of the patients were diagnosed as Depression. Both of these conditions are also the highest among the 278 conditions mentioned in the Annex II.

5.4 Establish and evaluate a rationalized sleep plan

(1) Reasonable sleep plan

Sleep is necessary to maintain the body's normal physiological function, and physical health and mental health are closely related. Sleep disorders can trigger a variety of adverse events, such as decreased quality of life, low work efficiency, traffic and industrial accidents, cognitive changes. Therefore, the quality of sleep has been paid more and more attention by people, Making a reasonable sleep plan plays an important role in ameliorating poor sleep quality and maintaining good sleep quality. Based on the results of the second and third question, we use Pittsburgh Sleep Quality Index (PSQI), which includes 7 items of subjective sleep quality, Sleep latency, sleep time, sleep efficiency, sleep disorders, Hypnagogue and Daytime dyfunction, A score of 0 to 3 points, the cumulative score of each item is the PSQI score, the higher the score, the worse the quality of sleep. PSQI score greater than or equal to 8 points as a standards for classificating good and poor sleep quality. Statistical analysis of the data in Annex II shows that most people who have poor quality of sleep account for the majority of the sample, and we formulate rationalized sleep plans for those with poor quality of sleep and those who want to continue to maintain good quality of sleep.

1) Rational sleep plan for people with poor sleep quality

According to a large number of sample statistical analysis of Annex II, the seven average scores of subjective sleep quality, Sleep latency, sleep time, sleep efficiency, sleep disorders, Hypnagogue and Daytime Dyfunction for poor quality of sleep was counted, The average score of Sleep latency , Subjective quality of sleep, Sleep efficiency, Sleep time, Sleep disorder, Daytime Dyfunction and Hypnagogue are respectively 2.676, 2.548, 2.425, 2.292, 1.871, 1.391, 1.330(See Table 5). So for the above phenomenon, poor quality of sleep staff should focus on minimizing the score of the four aspects in Sleep latency, Subjective quality of sleep, Sleep efficiency and Sleep time. Specific approach is made for improving the sleep quality of poor sleep conditions. Reading books before bedtime to shorten the time into falling asleep. Releasing pressure in all aspects of their lives and abandoning all kinds of ideological burden so as to improve the quality of subjective sleep and sleep efficiency. Sleeping early and getting up early to developing standard sleep schedule.

Table 5 PSQI score of various indicators of poor sleep quality

indicators	Sleep latency	Sub-Sleep quality	Sleep efficiency	Sleep time	Sleep disorder	Daytime dyfunction	Hypna-gogue
PSQI scores	2.676	2.548	2.425	2.292	1.871	1.391	1.330

2) Develop a reasonable sleep plan for the maintenance of good sleep quality

Through a large number of sample statistical analysis of Annex II, it was found that scores among the 7 items of subjective sleep quality, Sleep latency, sleep time, sleep efficiency, sleep disorders, Hypnagogue and Daytime dyfunction for those who continued to maintain good sleep quality , The average score of Sleep latency ,Subjective quality of sleep, Sleep

disorder, Sleep time, Sleep efficiency, Daytime Dyfunction and Hypnagogue are respectively 1.055, 1.022, 0.945, 0.674, the average of sleep time was 0.641, 0.599, 0.191.(See Table 6). In view of the above phenomenon, we should focus on four aspects in Subjective quality of sleep, Sleep disorder, Sleep time and Sleep efficiency, as much as possible to reduce the score of four scores. Specific approach is made so as to continue to maintain a good quality of sleep. Abandoning a variety of distractions before going to sleep, such as remove electronic devices. Keeping inner peace to shorten the time into falling asleep in order to improve the quality of subjective sleep during the day. Doing proper exercise and maintaining a good sleeping position.

Table 6 PSQI score of various indicators of good sleep quality

indicators	Sleep latency	Sub-Sleep quality	Sleep disorder	Sleep time	Sleep efficiency	Daytime dyfunction	Hypna-gogue
PSQI scores	1.055	1.002	0.945	0.641	0.599	0.674	0.191

(2) Sleep planning assessment

Utilizing analytic hierarchy process, 1-7 scale (as shown in Table 7) to build the judgment matrix.

Table 7 1-7 scale Meanings of a_{ij}

Scale a_{ij}	Explanation
1	The effects of C_i and C_j are the same
3	C_i slightly stronger than C_j
5	C_i is more influential than C_j
7	The effect of C_i is significantly stronger than that of C_j

Build a judgment matrix as follows:

$$A = \begin{bmatrix} 1 & 5 & 3 & 1 & 7 & 3 & 3 \\ \frac{1}{5} & 1 & \frac{1}{3} & \frac{1}{3} & 1 & 3 & \frac{1}{5} \\ \frac{1}{3} & 3 & 1 & 1 & 7 & 5 & 3 \\ 1 & 3 & 1 & 1 & 7 & 5 & 3 \\ \frac{1}{5} & 1 & \frac{1}{7} & \frac{1}{7} & 1 & 3 & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{5} & \frac{1}{5} & \frac{1}{3} & 1 & \frac{1}{3} \\ \frac{1}{3} & 5 & \frac{1}{3} & \frac{1}{3} & 5 & 3 & 1 \end{bmatrix}$$

Check the consistency of judgment matrix

Consistency check

$$CI = \frac{\lambda - n}{n - 1}$$

Where: λ is the largest eigenvalue of matrix A.,n=7.

calculated：λ=7.812, CI=0.135.

Table 8 Random consistency index RI value

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

As can be seen from Table 8, when n = 7, RI = 1.32, and when $n \geq 3, CR = \frac{CI}{RI} < 0.1$, it is considered that the degree of inconsistency of A is within the allowable range

Get the consistency ratio CR = 0.098 <0.1, through the consistency test, using Excel to get all the factorsFull-time coefficient:

Table 9 Full-time coefficient between indicators

indicators	subjective sleep quality	Sleep latency	Sleep time	Sleep efficiency	Sleep disorder	Hypna-gogue	Daytime dyfunction	weight coefficient
subjective sleep quality	1.000	5.000	3.000	1.000	7.000	3.000	3.000	0.245
Sleep latency	0.200	1.000	0.333	0.333	1.000	3.000	0.200	0.065
Sleep time	0.333	3.000	1.000	1.000	7.000	5.000	3.000	0.217
Sleep efficiency	1.000	3.000	1.000	1.000	7.000	5.000	3.000	0.224
Sleep disorder	0.200	1.000	0.143	0.143	1.000	3.000	0.200	0.061
Hypnagogue	0.333	0.333	0.200	0.200	0.333	1.000	0.333	0.029
Daytime dyfunction	0.333	5.000	0.333	0.333	5.000	3.000	1.000	0.160

It can be seen from Table 9 , subjective sleep quality, Sleep latency, Sleep time, Sleep efficiency, Sleep disorder, Hypnagogue and Daytime Dyfunction are expressed by w_1、w_2、w_3、w_4、w_5、w_6、w_7 respectively. Because of the different suggestions for people with different sleep conditions are different, for those with poor quality of sleep to consider Sleep latency, subjective sleep quality, sleep efficiency, sleep time,, so after taking recommendations to improve sleep quality as follows:

$$s_1 = \frac{w_1 + w_2 + w_3 + w_4}{\sum_{i=1}^{7} w_i} = 75.1\%$$

For those who continue to maintain a good quality of sleep People consider Sleep latency, subjective sleep , sleep disorders, daytime defunction, so taking suggestions, the quality of sleep to enhance results .

$$s_1 = \frac{w_1 + w_2 + w_5 + w_7}{\sum_{i=1}^{7} w_i} = 53.1\%$$

6. The model test

The calculated result of the theoretical model is obtained under steady-state conditions, and the rationality of the model needs to be further verified. In order to verify the rationality of the theoretical model, some data are randomly selected in Schedule 2, and the data are tested by a decision tree model to determine the symptom. Compare the result of the model prediction with the actual result, if the result predicted by the theoretical model is consistent with the actual result within a certain error range, the rationality and correctness of the theoretical model will be demonstrated. Table 10 below shows some of the results from the decision tree model prediction (see appendix 1 for a complete list).

Table 10 Part of the results predicted by the decision tree model

NO.	Sleep quality	Sleep latency	Sleep time	Sleep efficiency	Sleep disorder	Hypna -gogue	Daytime dyfunction	Diagnosis	Pre-
1	3	3	3	3	2	0	1	3	3
2	3	3	2	3	2	3	1	18	18
3	2	2	1	3	1	0	1	1	2
4	2	2	0	1	1	0	1	2	2
5	1	0	1	0	1	0	1	3	3
6	3	2	2	3	3	0	3	2	2
7	2	2	1	2	2	1	0	1	1
8	3	3	3	3	1	3	0	2	1
9	2	1	1	2	1	0	1	6	6
10	2	3	3	3	1	3	1	5	5

A total of 41 samples were randomly selected, and the predicted result was the same as the actual one with 30 and the accuracy rate was 73%. So the decision tree model was reasonable, considering the data and some accidental errors.

7. The model's evaluation

7.1 The advantages of the model

(1) The decision tree model has low requirements on the independence between indicators;

(2) Decision tree model does not need to worry about whether the discrete points and the data are linearly separable;

(3) The model is simple and clear. Easy to use mathematical tools, such as Python, etc., it reduces the difficulty of solving the programming, shorten the running time and improve work efficiency.

(4) The decision tree model is the most influential method in machine learning, which has the advantages of easy to explain, high recognition efficiency and the rules of discriminant.

7.2 shortcomings of the model

(1) Decision tree model is easy to overfit;

(2) The accuracy of the model is not well judged.

8. The promotion of the model

Decision tree models are usually used in data mining. The models have a very wide range of applications in different areas, such as credit scoring, precision marketing, fraud prevention and more.

Reference

[1] Tobias Kaufmann, The brain functional connectome is robustly altered by lack of sleep, Neuroimage 127: 324-332, 2016.

[2] Nicole A. short, Norman B.schmid, A multimethod examination of the effect of insomnia symptons on anxious responding to a social stressor, Behavior Therapy 7894:30119-3, 2017.

[3] Qi Yuan Jiang, Mathematical Models, BeiJing,higher education press,2011.

[4] Quinlan JR. Induction of decision trees. Machine Learning, 1986,1(1): 81-106

[5] Hastie T,Tibshirani R,Friedman JH. The Element of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag,2001.

Appendix

Program 1

```
% Use matlab to do gradually return
%J is imported
clear,clc
x=J(:,1:6);
y=J(:,7);
stepwise(x,y)

%this tool is matlab
clear,clc
%Consistency check on A matrix
A=[1 5 3 1 7 3 3;
    1/5 1 1/3 1/3 1 3 1/5;
    1/3 3 1 1 7 5 3;
    1 3 1 1 7 5 3;
    1/5 1 1/7 1/7 1 3 1/5;
    1/3 1/3 1/5 1/5 1/3 1 1/3;
    1/3 5 1/3 1/3 5 3 1];
[v,d]=eigs(A);
%The largest characteristic root
ftmax=max(d(:))
%Get eigenvector rows and columns
[m,q]=size(v);
%The eigenvector is normalized
sum=0;
for i=1:q
sum = sum + v(i,1);
end
%tbvector = v(:,1);
for i=1:m
ftvector(i,1)= v(i,1)/sum;
end
%ftvector
n=7;
CI=(ftmax-n)/(n-1)
switch (1<=n&n<=15)
case n==1
p=0;
case n==2
p=0;
case n==3
p=0.52;
case n==4
```

```
p=0.89;
case n==5
p=1.12;
case n==6
p=1.26;
case n==7
p=1.36;
case n==8
p=1.41;
case n==9
p=1.46;
case n==10
p=1.49
case n==11
p=1.52
case n==12
p=1.54
case n==13
p=1.56;
case n==14
p=1.58;
case n==15
p=1.59;
otherwise p='error';
end
CR=CI/p
```

Program 2 Python: Decision tree program

```
# -*- coding:utf-8 -*-
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
import csv


if __name__ == '__main__':
    path = 'Annex_II.csv'
```

```
    data = pd.read_csv(path)
    data1 = pd.read_csv('Annex_III.csv')
    x = data[['Sleep quality', 'Sleep latency', 'Sleep time','Sleep efficiency','Sleep
disorder','Hypnagogue','Daytime dyfunction']]
    x1 = data[['Sleep quality', 'Sleep latency', 'Sleep time', 'Sleep efficiency', 'Sleep disorder',
'Hypnagogue','Daytime dyfunction']]
    y = data['Diagnosis1']
    x = np.array(x)
    y = np.array(y)
    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=1)
    model = Pipeline([
        ('ss', StandardScaler()),
        ('DTC', DecisionTreeClassifier(criterion='entropy', max_depth=6))])
    model = model.fit(x_train, y_train)
    y_test_hat = model.predict(x_test)
    y_test_hat1 = model.predict(x1)

    csvfile = open('csv.csv', 'wb')
    writer = csv.writer(csvfile)
    writer.writerows(x_test[:,:100])
    csvfile.close()

    # save
    # dot -Tpng -o 1.png 1.dot
    f = open('.\\MathBuild_2.dot', 'w')
    tree.export_graphviz(model.get_params('DTC')['DTC'], out_file=f)

    result = (y_test_hat == y_test)
    acc = np.mean(result)*2
    print 'Correct rate: %.2f%%' % (100 * acc)

    depth = np.arange(1, 15)
    acc_list = []
    for d in depth:
        clf = DecisionTreeClassifier(criterion='entropy', max_depth=d)
        clf = clf.fit(x_train, y_train)
        y_test_hat = clf.predict(x_test)
        result = (y_test_hat == y_test)
        acc =    np.mean(result)*2
        acc_list.append(acc)
        print d, 'Correct rate: %.2f%%' % (100 * acc)
    plt.figure(facecolor='w')
    plt.plot(depth, acc_list, 'ro-', lw=2)
    plt.xlabel(u'Depth of DT', fontsize=15)
```

```
plt.ylabel(u'Correct rate', fontsize=15)
plt.title(u'Depth of DT & Over Fitting', fontsize=17)
plt.grid(True)
plt.show()
```

Table 1. The prediction results of the decision tree model

NO.	Sleep quality	Sleep latency	Sleep time	Sleep efficiency	Sleep disorder	Hypnagogue	Daytime dyfunction	Diagnosis	Pre-
1	3	3	3	3	2	0	1	3	3
2	3	3	2	3	2	3	1	18	18
3	2	2	1	3	1	0	1	1	2
4	2	2	0	1	1	0	1	2	2
5	1	0	1	0	1	0	1	3	3
6	3	2	2	3	3	0	3	2	2
7	2	2	1	2	2	1	0	1	1
8	3	3	3	3	1	3	0	2	1
9	2	1	1	2	1	0	1	6	6
10	2	3	3	3	1	3	1	5	5
11	3	3	2	1	2	0	1	2	2
12	2	1	2	3	1	0	0	23	1
13	3	3	3	2	2	3	3	2	2
14	3	3	2	3	2	0	2	1	1
15	2	3	0	2	2	0	2	1	1
16	2	3	3	3	3	0	1	1	1
17	1	0	0	0	1	0	0	1	2
18	1	3	2	3	1	3	0	2	1
19	2	3	1	2	2	3	2	2	2
20	2	3	3	2	2	3	3	2	2
21	3	3	3	3	1	3	3	1	1
22	2	3	3	3	3	0	2	5	5
23	2	3	3	3	1	0	1	37	1
24	2	1	1	1	1	0	0	7	7
25	3	3	3	3	2	3	1	3	3
26	3	3	3	3	2	3	1	1	3
27	3	3	3	3	2	2	2	1	1
28	3	2	3	3	2	3	3	2	2
29	2	3	3	2	2	3	1	3	3
30	3	3	3	3	2	0	0	1	1
31	2	3	2	2	1	0	1	3	2
32	1	0	1	0	2	0	1	2	2
33	1	3	1	2	2	3	1	1	1
34	2	3	3	3	1	3	3	170	1
35	2	2	2	1	2	2	0	4	4

NO.	Sleep quality	Sleep latency	Sleep time	Sleep efficiency	Sleep disorder	Hypnagogue	Daytime dyfunction	Diagnosis	Pre-
36	2	3	1	3	2	1	1	5	5
37	2	3	2	3	2	1	0	108	1
38	3	3	0	1	2	0	0	1	1
39	2	3	3	3	2	0	2	3	3
40	3	3	3	3	3	1	2	3	2
41	2	3	0	1	1	0	2	4	4