I Data Preprocessing

1. Data cleaning, data integration

Since the computer test, online learning, and enrollment information data all use sid as a symbol to mark a student, so I choose to store the three types of information in the database, the data can be associated through the natural connection of sql and the select statements.

(1) Computer test data

The useful data in the computer-based test data are sid and ranking. Other information, such as the total time spent on answering questions, is related to the number and difficulty of test questions, which cannot directly reflect the degree of knowledge mastery of students, so they are discarded. By observing the data, it can be found that there are 4 normal exams and 1 final exam, which are divided into three classes, so five tables are used for storage, and each table has only sid and rank fields.

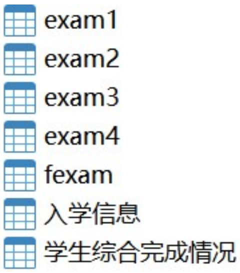
(2) Enrollment information data

There is only one xlsx file for the enrollment information data, so it can be stored in the database as it is, and the field names should be consistent with the original file

(3) Online learning data

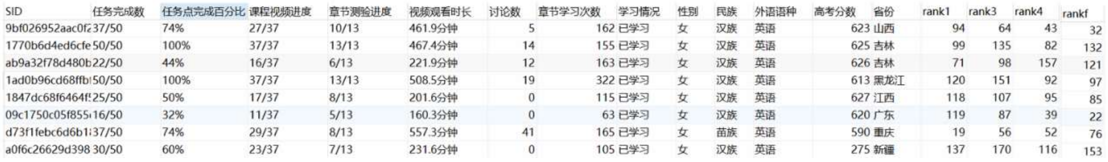
The online learning data selects the comprehensive completion status of students and stores it in the database as it is, with the field names consistent with the original file

Then through the above processing, there are seven tables in the database:



- exam1
- exam2
- exam3
- exam4
- fexam
- 入学信息
- 学生综合完成情况

At this time, all the data corresponding to a sid can be obtained by using the natural connection, and the data rows with missing data can be automatically filtered due to the characteristics of the natural join. By executing the count(*) statement, it can be found that the number of participants in the second test is only 399 , while the number of other exams is 499, and because the second exam has relatively less value than the first exam and the fourth exam, the second exam is omitted and not considered, and finally 493 rows data are obtained, the following data can be obtained at this time:



SID	任务完成数	任务点完成百分比	课程视频进度	章节测验进度	视频观看时长	讨论数	章节学习次数	学习情况	性别	民族	外语语种	高考分数	省份	rank1	rank3	rank4	rankf
9bf026952aac0f	37/50	74%	27/37	10/13	461.9分钟	5	162	已学习	女	汉族	英语	623	山西	94	64	43	32
1770b6d4ed6cfe	50/50	100%	37/37	13/13	467.4分钟	14	155	已学习	女	汉族	英语	625	吉林	99	135	82	132
ab9a32f78d480b	22/50	44%	16/37	6/13	221.9分钟	12	163	已学习	女	汉族	英语	626	吉林	71	98	157	121
1ad0b96cd68ffb	50/50	100%	37/37	13/13	508.5分钟	19	322	已学习	女	汉族	英语	613	黑龙江	120	151	92	97
1847dc68f6464f	25/50	50%	17/37	8/13	201.6分钟	0	115	已学习	女	汉族	英语	627	江西	118	107	95	85
09c1750c05f855	16/50	32%	11/37	5/13	160.3分钟	0	63	已学习	女	汉族	英语	620	广东	119	87	39	22
d73f1febc6d6b1	37/50	74%	29/37	8/13	557.3分钟	41	165	已学习	女	苗族	英语	590	重庆	19	56	52	76
a0f6c26629d398	30/50	60%	23/37	7/13	231.6分钟	0	105	已学习	女	汉族	英语	275	新疆	137	170	116	153

2. Data Transform

(1) SID

To make sid more intuitive, use the python dictionary to map to 0~492.

(2) The number of tasks completed and the percentage of tasks completed

The task completion percentage is the percentage of the task completion number, so only the task completion number can be taken.

(3) Video viewing time

Since the video viewing time data is evenly distributed, it can be processed directly using max-min normalization.

(4) Number of discussions

Since the maximum number of discussions is 220 and the second largest number of discussions is 56, the difference between the two is large, so the maximum value is discarded, and the second largest number of discussions is taken as the maximum value for maximum-minimum normalization.

(5) Number of chapters studied

Since the maximum number of chapters learning times is 652 and the second largest is 546 times, the difference between the two is large, so the maximum value is discarded, and the second largest chapter learning times are taken as the maximum value for maximum-minimum normalization.

(6) College Entrance Examination Scores

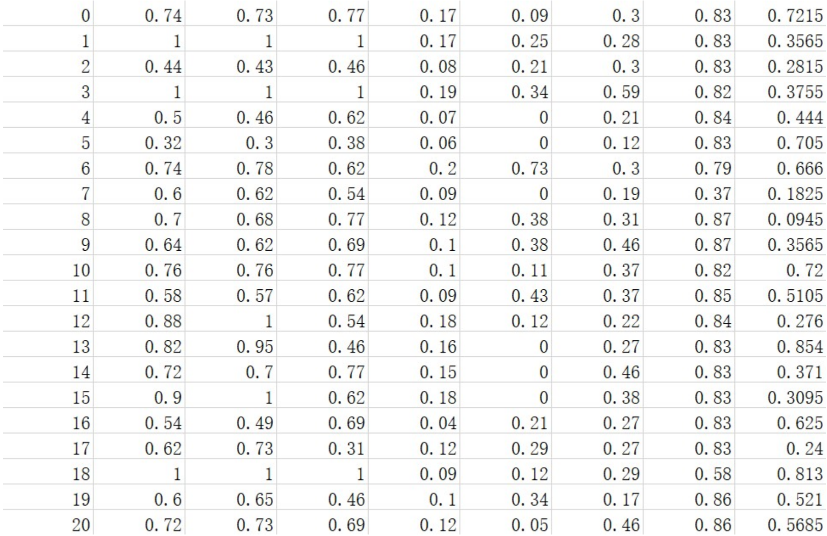
Since the total score of the college entrance examination is related to the province, the total score of the college entrance examination is obtained through the province, and then the score of the college entrance examination is divided by the total score of the college entrance examination for data normalization.

(7) Ranking

Since the second test was omitted, there are four rankings for the first, third, fourth, and final exams. The maximum number of people in the three classes is 173 as the maximum value for normalization, and the processing result is set to rank1, rank3, rank4, rankf, and then use the following formula to calculate the total score:

$$\text{Total Score} = 1 - (\text{rank1}\times 0.1+\text{rank3}\times 0.2+\text{rank4}\times 0.35+\text{rankf}\times 0.35)$$

After the above processing, the following data is finally obtained and saved as an .xls file:



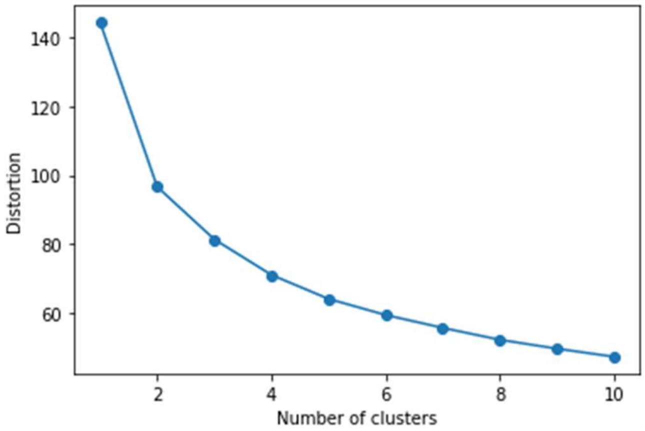
0	0.74	0.73	0.77	0.17	0.09	0.3	0.83	0.7215
1	1	1	1	0.17	0.25	0.28	0.83	0.3565
2	0.44	0.43	0.46	0.08	0.21	0.3	0.83	0.2815
3	1	1	1	0.19	0.34	0.59	0.82	0.3755
4	0.5	0.46	0.62	0.07	0	0.21	0.84	0.444
5	0.32	0.3	0.38	0.06	0	0.12	0.83	0.705
6	0.74	0.78	0.62	0.2	0.73	0.3	0.79	0.666
7	0.6	0.62	0.54	0.09	0	0.19	0.37	0.1825
8	0.7	0.68	0.77	0.12	0.38	0.31	0.87	0.0945
9	0.64	0.62	0.69	0.1	0.38	0.46	0.87	0.3565
10	0.76	0.76	0.77	0.1	0.11	0.37	0.82	0.72
11	0.58	0.57	0.62	0.09	0.43	0.37	0.85	0.5105
12	0.88	1	0.54	0.18	0.12	0.22	0.84	0.276
13	0.82	0.95	0.46	0.16	0	0.27	0.83	0.854
14	0.72	0.7	0.77	0.15	0	0.46	0.83	0.371
15	0.9	1	0.62	0.18	0	0.38	0.83	0.3095
16	0.54	0.49	0.69	0.04	0.21	0.27	0.83	0.625
17	0.62	0.73	0.31	0.12	0.29	0.27	0.83	0.24
18	1	1	1	0.09	0.12	0.29	0.58	0.813
19	0.6	0.65	0.46	0.1	0.34	0.17	0.86	0.521
20	0.72	0.73	0.69	0.12	0.05	0.46	0.86	0.5685

Through observation, we can found that the data distribution is uniform, and there is no situation that generally approaches 0, so the processing result is satisfactory.

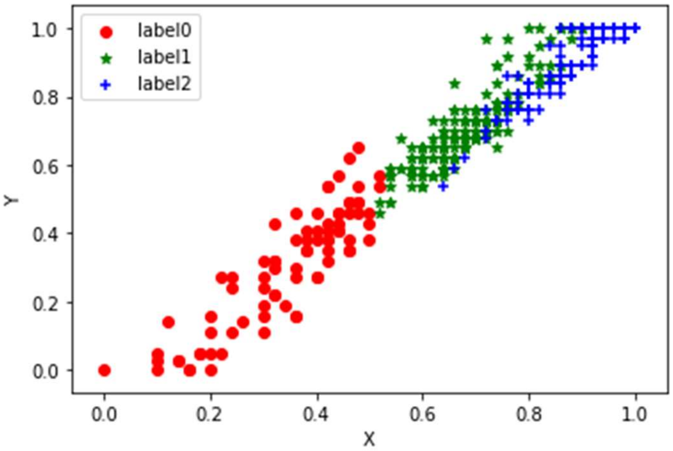
Ⅱ Data Merge

1. Euclidean distance

After data preprocessing, the meanings of each column of data are: SID, number of tasks completed, course video progress, chapter test progress, video viewing time, number of discussions, chapter learning times, college entrance examination scores, and total scores, so the Euclidean distance is adopted As a distance function, the K-Means algorithm is used to cluster all column information except SID, and the elbow method is used to obtain the following figure:



Therefore, we can see that it should be divided into three categories. After clustering, the following results are obtained, where X indicates the number of tasks completed, and Y indicates the progress of the course video:



The cluster centers are:

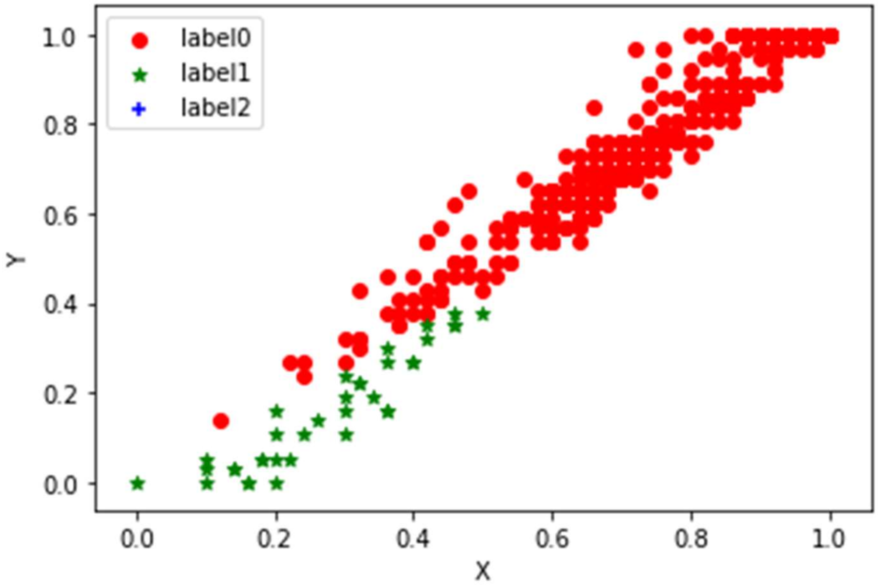
0	0.34375	0.304	0.457875	0.05975	0.08625	0.23	0.8115	0.4234875
1	0.69060465	0.70967442	0.63669767	0.13306977	0.16581395	0.34697674	0.82660465	0.46019767
2	0.91653061	0.9369898	0.85770408	0.21959184	0.25234694	0.48280612	0.82428571	0.61024745

Each column represents: category, number of tasks completed, course video progress, chapter test progress, video viewing time, number of discussions, chapter learning times, college entrance examination scores, and total scores.

From the above results, we can see that the value of each column in the three categories is generally category 2>category 1>category 0, indicating that the students can be divided into three categories by clustering: students with high learning enthusiasm, students with medium learning enthusiasm and students with low learning enthusiasm. Students with high learning enthusiasm can complete all the learning tasks better, so they have a higher total score in the end, while students with medium and low learning enthusiasm have lower total scores, and from the above results, we can see that the three types of students have similar college entrance examination scores, indicating that the college entrance examination score does not significantly affect the final total score of the course.

2. Cosine similarity

When the distance function is defined as cosine similarity, assuming there are students A, B, and C, the ratios of the number of tasks completed and the number of discussions of the three students are (0.3, 0.3), (0.9, 0.9), (0.6, 0.3), Then students A and B will be in the same cluster, while student C will be in a cluster different from A and B, which means that students A and B think that the number of completed tasks is as important as the progress of the course video, while student C thinks that the task completions are more important than discussions. The college entrance examination scores and total scores are not considered when clustering by cosine similarity. The significance of clustering is to cluster according to the importance of different learning activities. Set the category to 2 categories, and get the following results after clustering, where X indicates the number of tasks completed, and Y indicates the progress of the course video:

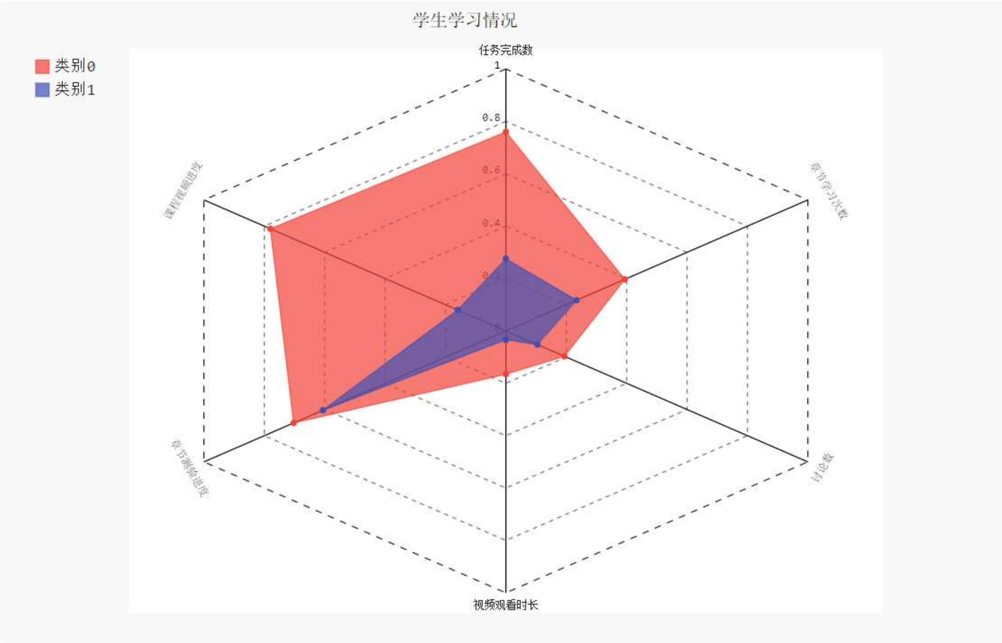


The cluster center points are:

0	0.75978022	0.77978022	0.7029011	0.1652967	0.19395604	0.39384615
1	0.27555556	0.15972222	0.60583333	0.03388889	0.10444444	0.23416667

Each column represents: category, number of tasks completed, course video progress, chapter test progress, video viewing time, number of discussions, chapter learning times.

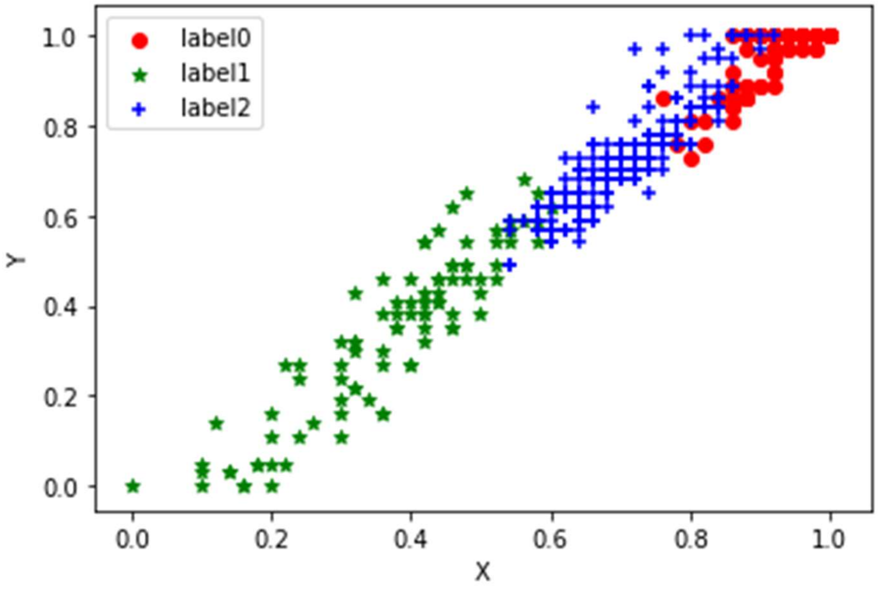
The radar chart can be drawn as follows:



From the figure above, we can see that students in category 0 tend to complete all learning tasks in a balanced manner, while students in category 1 only pay more attention to the progress of the chapter test.

3. Manhattan distance

Manhattan distance and Euclidean distance have similar functions in cluster analysis. After clustering, the following results are obtained, where X represents the number of tasks completed, and Y represents the course video progress:



The cluster center points are:

0	0.37376344	0.34225806	0.46408602	0.0655914	0.08387097	0.23526882	0.80924731	0.4241129
1	0.71138075	0.7276569	0.66556485	0.14870293	0.16941423	0.37246862	0.83138075	0.50297071
2	0.94867925	0.9736478	0.87672956	0.21880503	0.27496855	0.48257862	0.8191195	0.58350629

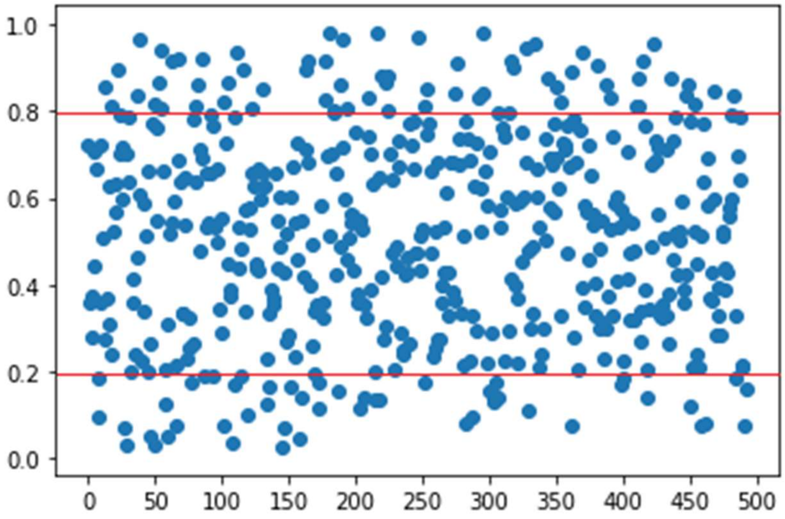
Each column represents: category, number of tasks completed, course video progress, chapter test progress, video viewing time, number of discussions, chapter learning times, college entrance examination scores, and total scores.

From the above chart we can see that the clustering results of Manhattan distance and Euclidean distance are similar.

Ⅲ Score Forecast

1. Forecast method

Use the data obtained in the data preprocessing stage, set the sid as the X axis, and set the student scores as the Y axis to draw a scatter diagram as follows:



From the above figure we can see that the distribution of student grades is concentrated between 0.2 and 0.8, so students whose grades are lower than 0.2 are regarded as low, students whose grades are between 0.2 and 0.8 are regarded as medium, and students whose grades are higher than 0.8 are regarded as high , and set the value to 0,1,2. Since the existing data is the data of the whole semester, and the maximum-minimum normalization is used in the preprocessing process, the data obtained after processing is relative data, and since the learning attitude of the students is stable, the data is basically equivalent to the data at the end of the semester after normalization. Therefore when the course is halfway through, the student data can be processed in the same way and input into the model to obtain a reference value. In summary, the model construction process is as follows:

1. Map the grade column of the dataset to 0,1,2

2. Divide the data set into a training set and a test set with a ratio of 7:3

3. Use the training set to build a random forest model

4. Use the test set to check the accuracy of the model

After using python to build the model according to the above process, the final accuracy of the model is 0.7905405405405406.

2. Forecast principle

Prediction can be divided into regression and classification. If the regression method is used, the student's grades are closely related to the student's online learning situation, but the student's grades still depend on the offline learning situation, so the classification method is selected for prediction. The random forest is composed of multiple decision trees, so it has a stronger classification ability than a single decision tree, so the random forest algorithm is selected for classification prediction.

3. Model evaluation

After using the training set to build the model, the accuracy rate of the test set is 0.79, which is relatively acceptable in terms of value. However, the division of grades in this model is based on a hierarchical system, and it is impossible to predict the value of students' grades, and the data used for training is the data at the end of the semester, which still has a certain range of deviation compared with the data in the middle of the semester, so there is still room for improvement.