**Design and implementation of SVM-based breast cancer dataset classification**

**Summary：**This report examines the use of machine learning algorithms to automate diagnosis with breast lump detection data. There were a total of 569 samples in the dataset, each containing 30 features of breast cells. These 30 features are composed of the mean, standard deviation and maximum values of the 10 basic characteristics of the digitized nucleus (including radius, texture, perimeter, area, symmetry, etc.). Using a support vector machine, our team constructed a classifier for breast cancer diagnosis. After training and testing, the accuracy of this model reaches 96.491%.

**Keywords**: breast cancer, machine learning, classification model, support vector machine

1 Introduction

Breast cancer is characterized by high complication rates and high mortality, which seriously threatens the health of women worldwide.

According to the World Health Organization (WHO), there were 2.09 million breast cancer cases worldwide in 2018, of which 627,000 were deaths. Breast cancer has become the most common malignant tumor and the second leading cause of cancer death in women worldwide.

According to the American Cancer Society, breast cancer mortality rates have been reduced by 40 percent between 1989 and 2016. Since 2007, the mortality rate for women under 50 has not decreased significantly, but it continues to decline for women over 50.

These advances are thanks to better treatments and early diagnosis of breast cancer.

Currently, the molecular typing of breast cancer needs to be determined by needle biopsy. This experiment uses machine learning technology to establish a classifier based on the characteristic data related to breast cancer molecular typing to assist in breast cancer prediction and early diagnosis.

2 Formulation of questions, introduction of methods and models

The aim of this study is to build a classifier with the highest possible accuracy using a classification model.

There are 2 categories in this dataset, namely "benign" and "malignant", and 30 cell characteristics. According to the characteristics of the dataset, we select the support vector machine as the classification model.

Support vector machines (SVMs) are binary classification models, the purpose of SVMs is to find a line to "best" distinguish between these two types of points, so that if there are new points in the future, this line can also be well classified, which is illustrated in two dimensions. In a high-dimensional space, we want to distinguish between two types of sample data, and we need to find a hyperplane to distinguish between the two types of sample data. SVMs are suitable for small and medium-sized data samples, nonlinear, high-dimensional classification problems.

"38 lines" can be seen as an image interpretation of SVM in two-dimensional space, which conveys the following important information:

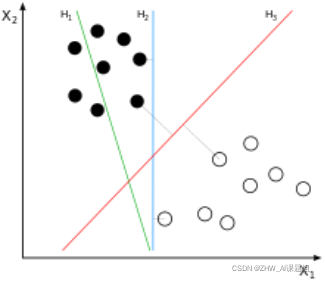
(1) is a straight line (linear function);

(2) Can divide the desktop into two parts, which belong to you and me respectively (with classification function, which is a binary classification);

(3) Located in the middle of the desk, not biased to either side (focusing on the principle of fairness can ensure the maximum interests of both parties).

The above three points are the central idea of the SVM algorithm.

SVM algorithm interpretation and analysis: SVM will look for a division hyperplane that can distinguish between two classes and maximize the margin. A relatively good division hyperplane has the least influence on the sample local disturbance, produces the best classification results, and has the strongest generalization ability for unseen examples. As can be seen from the figure below, H1 is linearly indivisible, and H2 and H3 are linearly separable. At this time, we use the principle of the largest interval to select H3 as the hyperplane to distinguish the two types of sample points as the figure below.

****Figure.SVM hyperplane

In fact, most of the time the data is not linearly separable, and a hyperplane that satisfies such conditions does not exist at all. For nonlinear cases, SVM is processed by selecting a kernel function κ(.,.), and finally constructing an optimal separation hyperplane in the high-dimensional feature space by mapping the data to the high-dimensional feature space, thereby separating the nonlinear data on the plane itself.

3 Experimental principle and steps

3.1 Experimental principle

Note that xi (i = 1, 2...30) is 30 characteristic variables, [ai,yi] (i=1,2...569) is the sample, ai belongs to a 30-dimensional linear space, benign yi = -1, malignant yi = -1.

Define the optimal division (w\*X)+b=0,

X=[x1,x2,... x30], w belongs to a 30-dimensional space, satisfied

(w\*ai)+b>=1,yi=1

(w\*ai)+b<=-1,yi=-1

where, the equation is satisfied

The sample of (w\*ai)+b=+-1 is a support vector.

To maximize the distance between the classification polygons of the two types of populations, there is

Max 2/|w|

namely

Min |w|^2/2

Therefore, the following mathematical model of SVM is established:

Min |w|^2/2

s.t. yi\*[(w\*ai)+b]>=1

(i=1,2,... 30)

Finding the w and b corresponding to the optimal values yields the classification function

g(x)=sign[(w\*X)+b]

3.2 Experimental steps

3.2.1 Datasets

The experimental dataset for this study is from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. These features were calculated from digitized images of a fine needle aspiration (FNA) breast mass. They describe the characteristics of the nucleus present in the image.

Number of instances: 569

Number of attributes: 32 (ID, diagnostics, 30 real-valued features)

Property information

1) ID number

2) Diagnosis (M = malignant, B = benign)

Real-valued characteristics:

a) Radius

b) Textures

c) Perimeter

d) Area

e) Smoothness

f) Compactness

g) Concaveness

h) Pit points

i) Symmetry

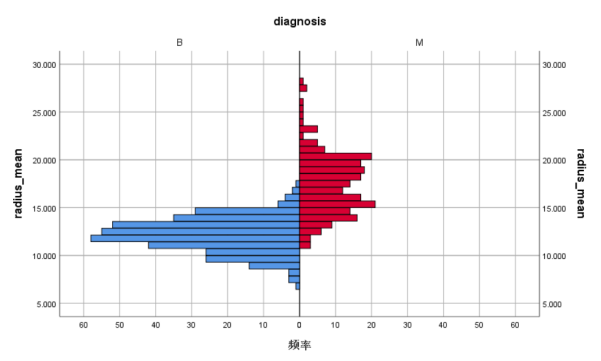
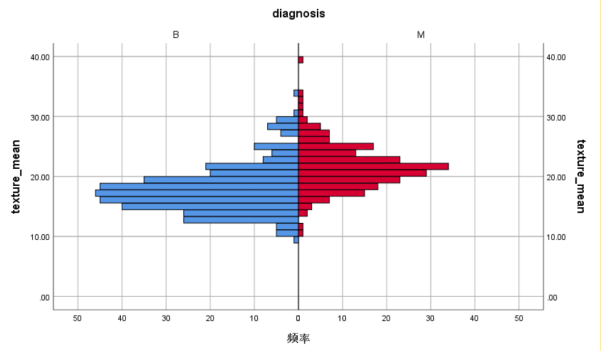
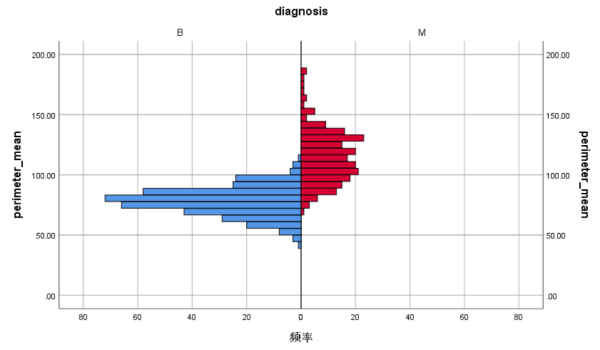
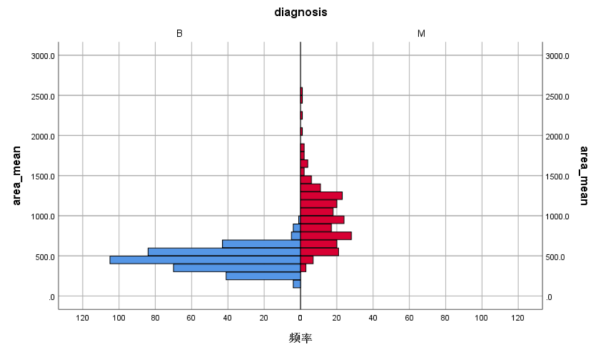
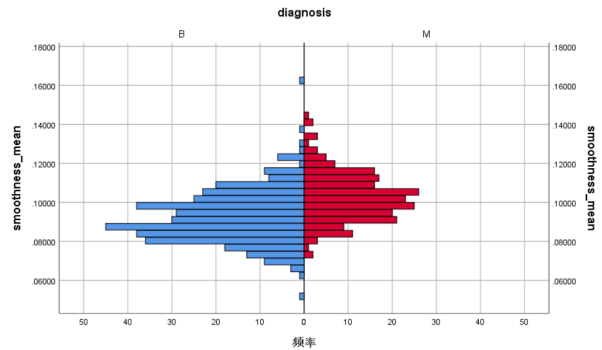
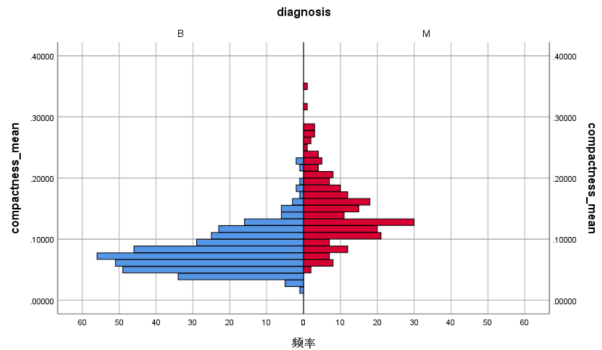
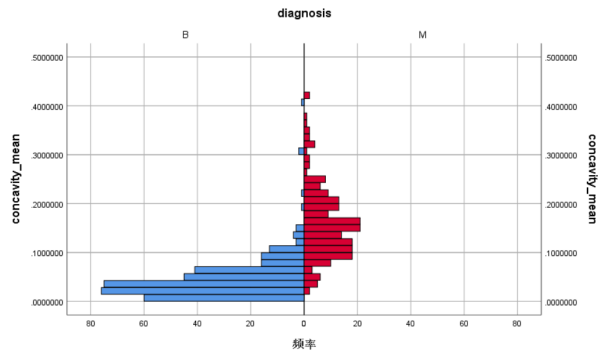
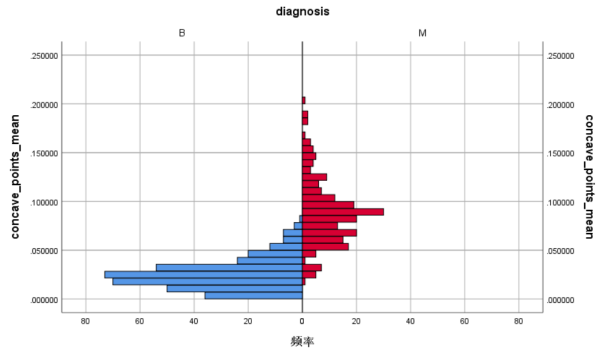
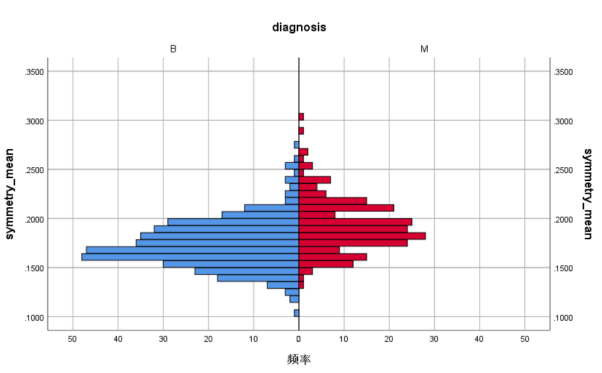
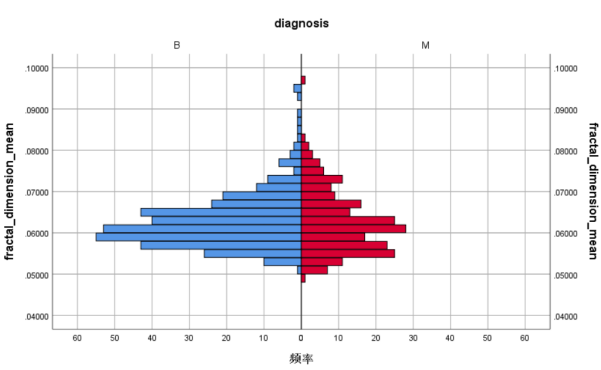
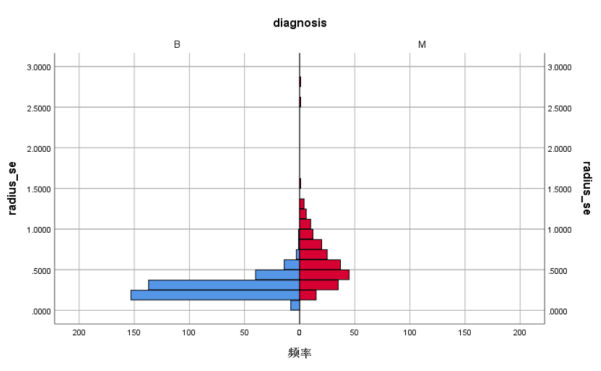
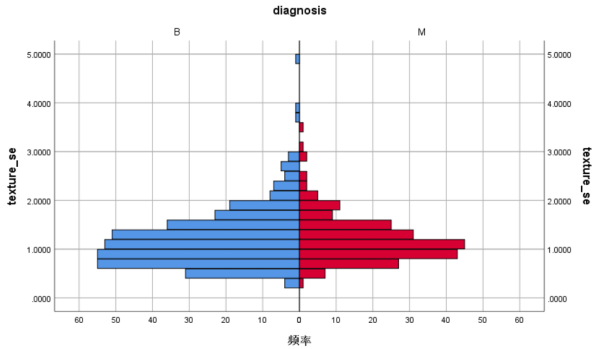
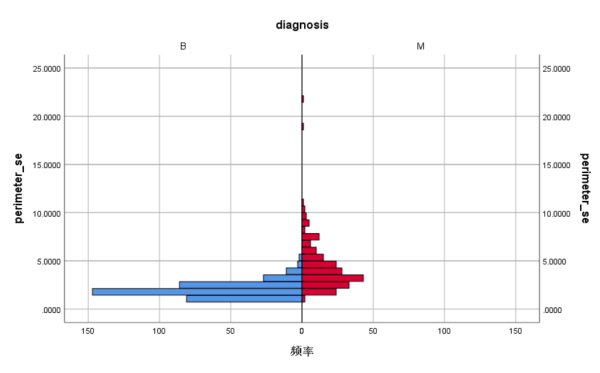
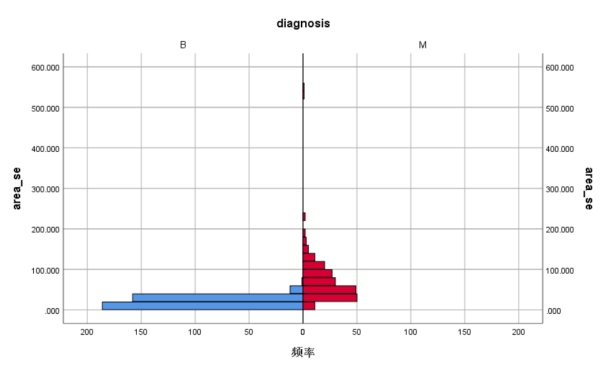
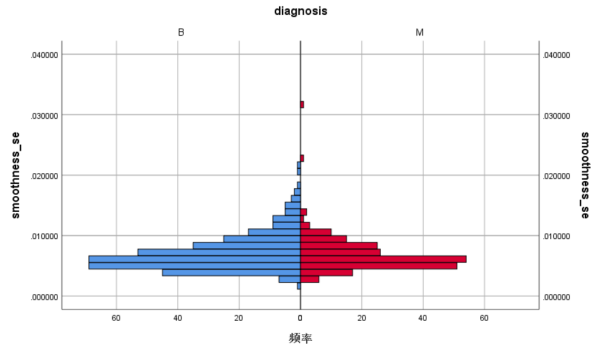
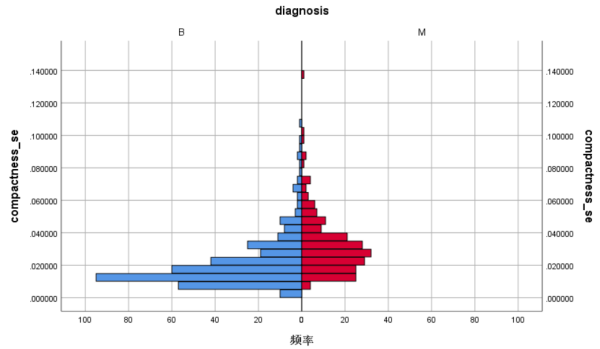
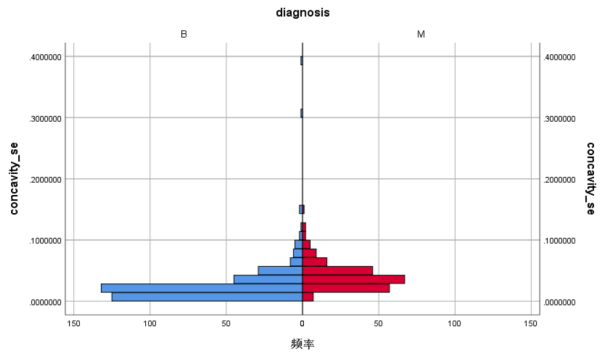
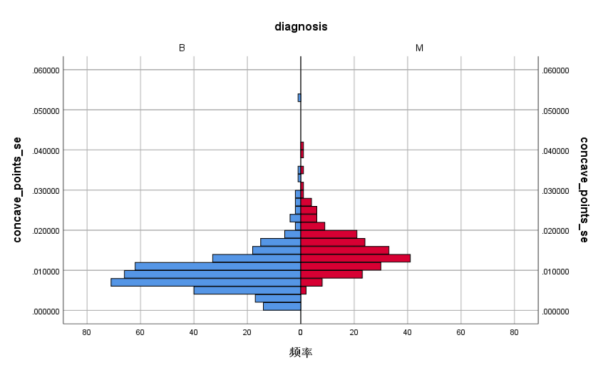
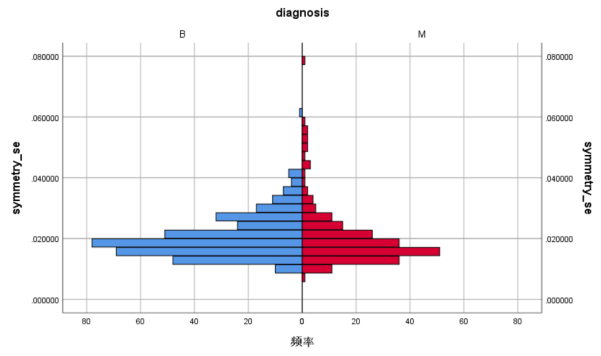
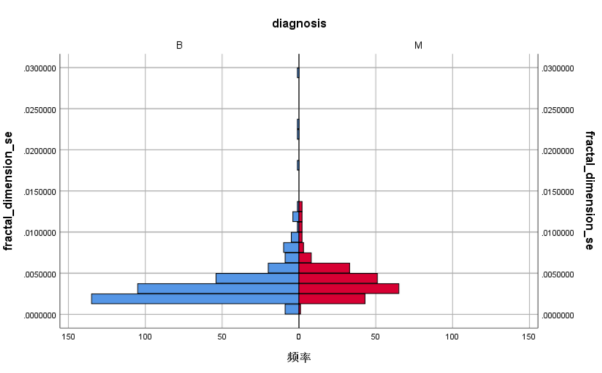
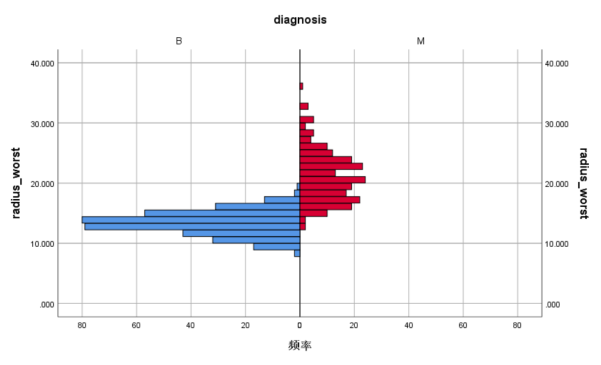
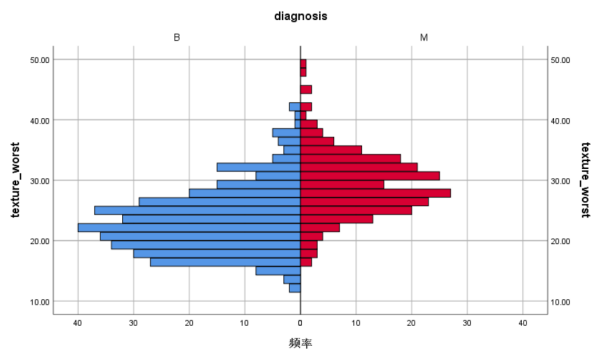
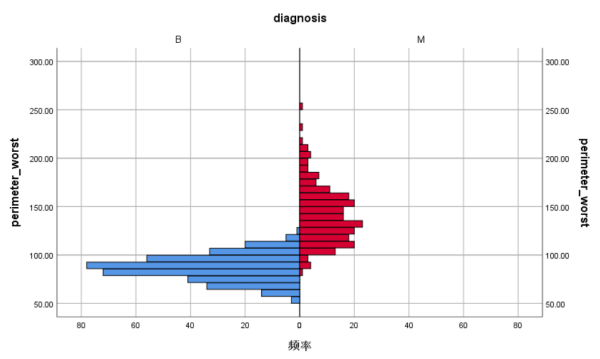
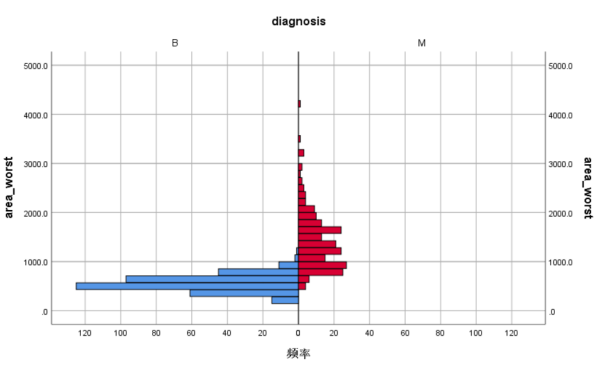
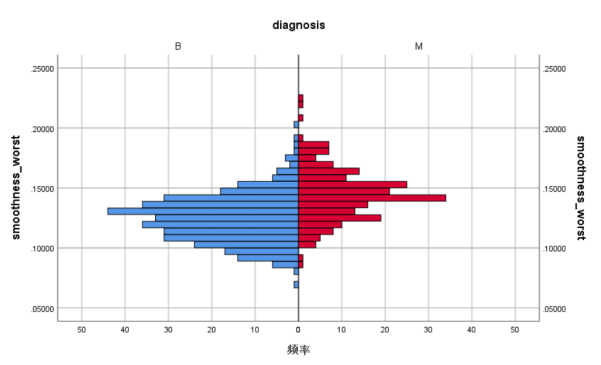
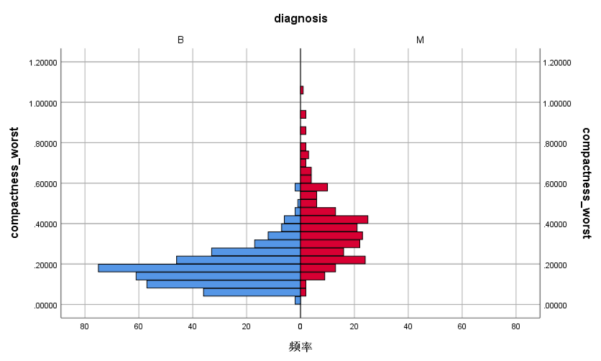
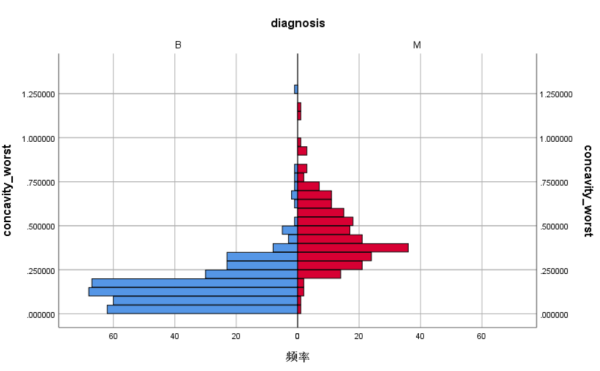
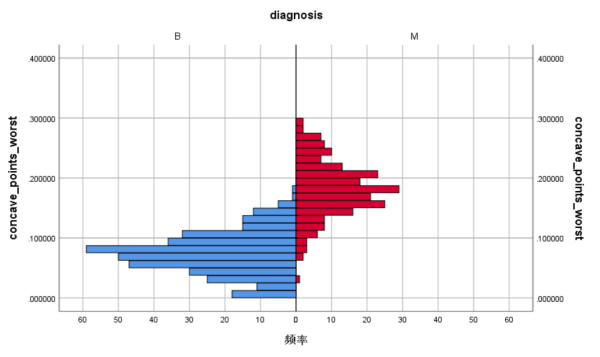
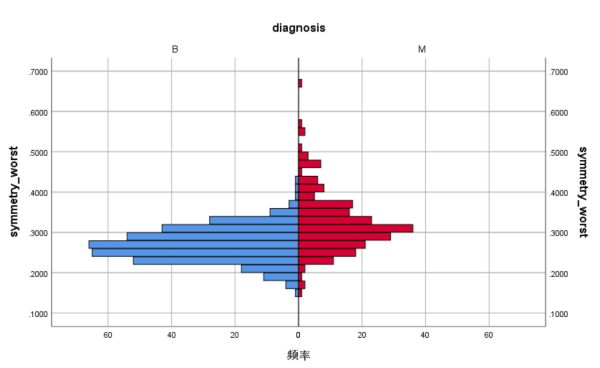
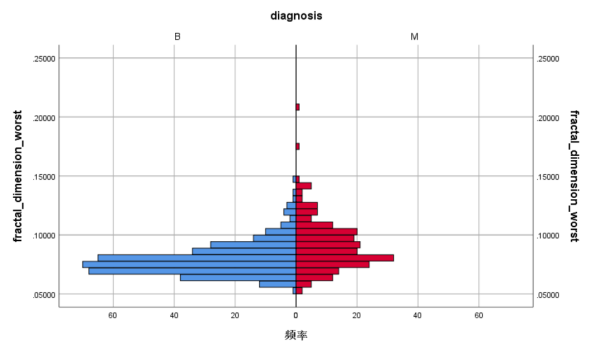
j) fractal dimension

Mean, standard deviation, and maximum (worst) of the above real-valued characteristics

We convert the provided data file into a csv file and add the feature name to the first line.

3.2.2 Data preprocessing

First, the characteristics in the dataset are analyzed by the Population Pyramid chart.

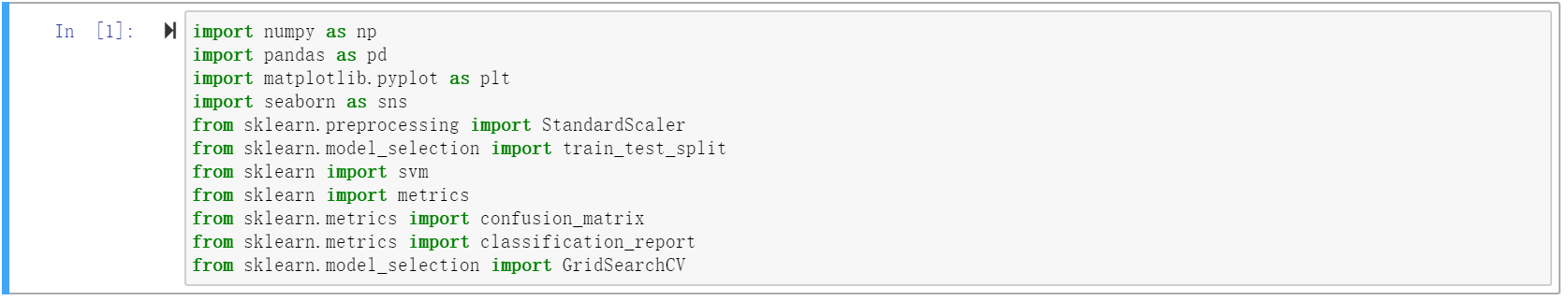
1. Mean radius
2. Texture mean
3. Perimeter mean
4. The mean area
5. Smoothness mean
6. Mean compactness
7. Mean concavity
8. Concave points mean
9. Mean symmetry
10. Mean fractal dimensio
11. Radius standard deviation 
12. Texture standard deviation
13. Perimeter standard deviation
14. Standard deviation of area
15. Standard deviation of smoothness
16. Standard deviation of compactness
17. Concavity standard deviation
18. Concave standard deviation
19. Standard deviation of symmetry
20. Standard deviation of fractal dimension
21. Maximum radius
22. The maximum texture value
23. Perimeter maximum
24. The maximum area
25. The maximum smoothness degree
26. Maximum compactness
27. Maximum concavity
28. The maximum concave point
29. Maximum symmetry
30. The maximum fractal dimension

Through the analysis of the distinction between classes of the above 30 features. It is not difficult to find that the radius, perimeter, area, concave point and concave degree are more obvious, and the mean discrimination degree of each attribute is also high. The order of feature importance is as follows:

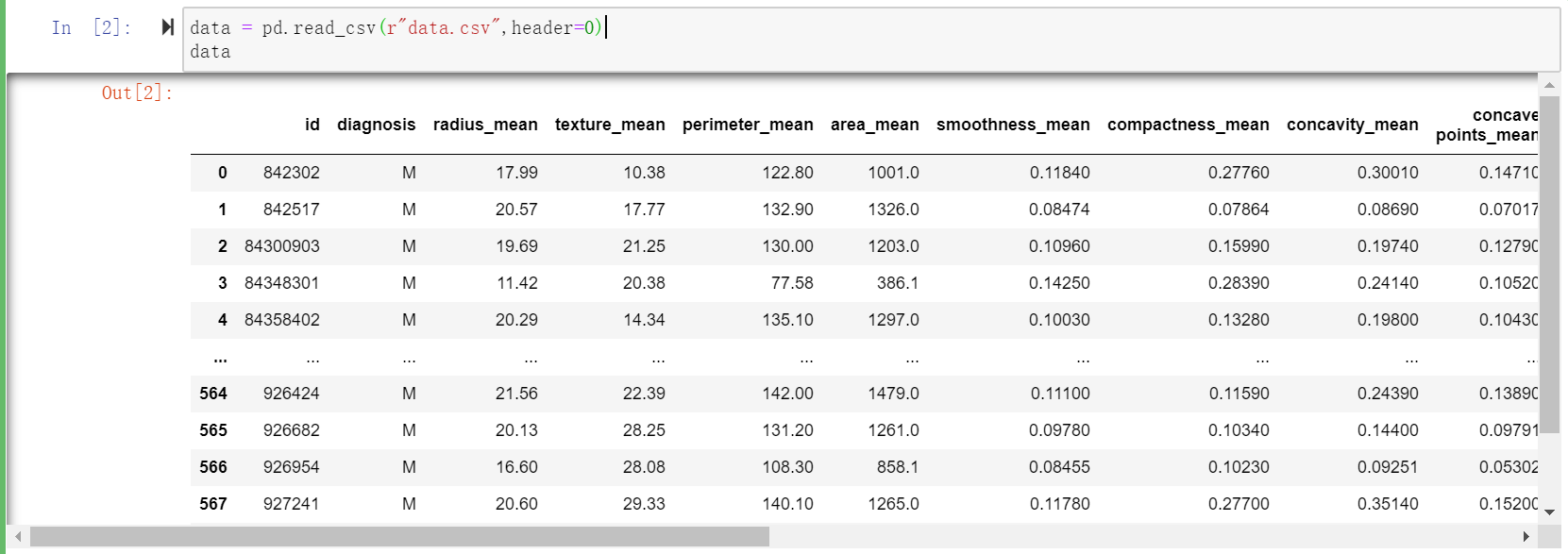
Mean radius = mean perimeter = mean area > mean pit = mean concave point

3.3.3Preprocess the code

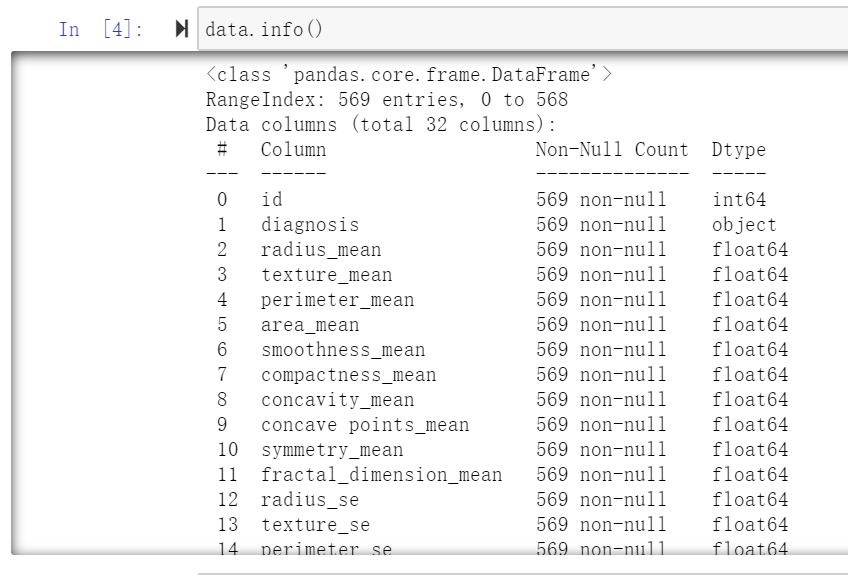
Import the libraries you want to use



Open the dataset



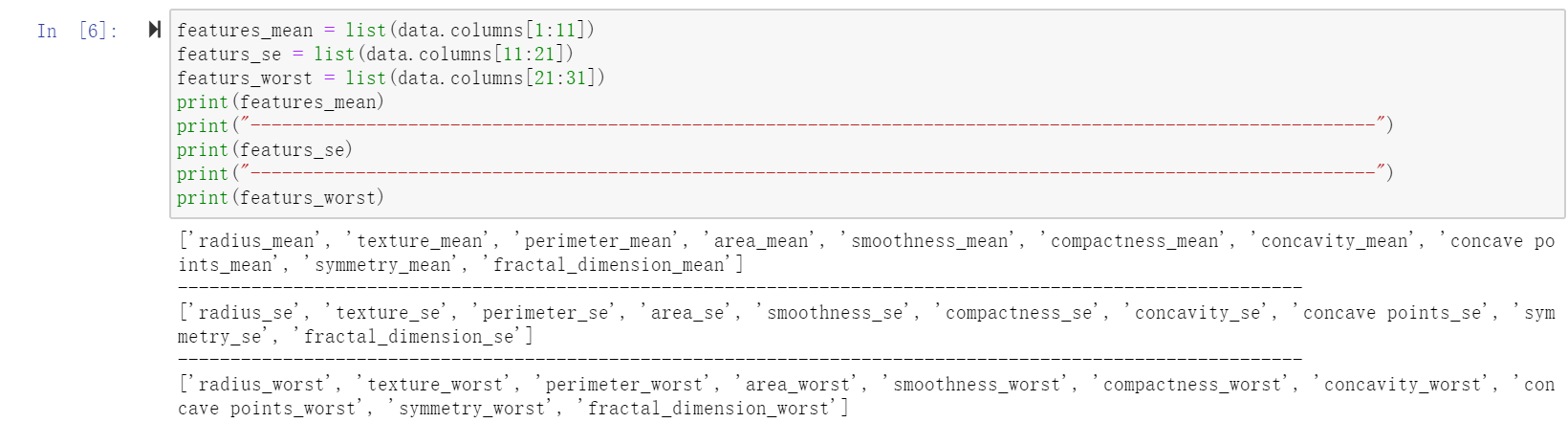
Check the feature values and data types in the dataset



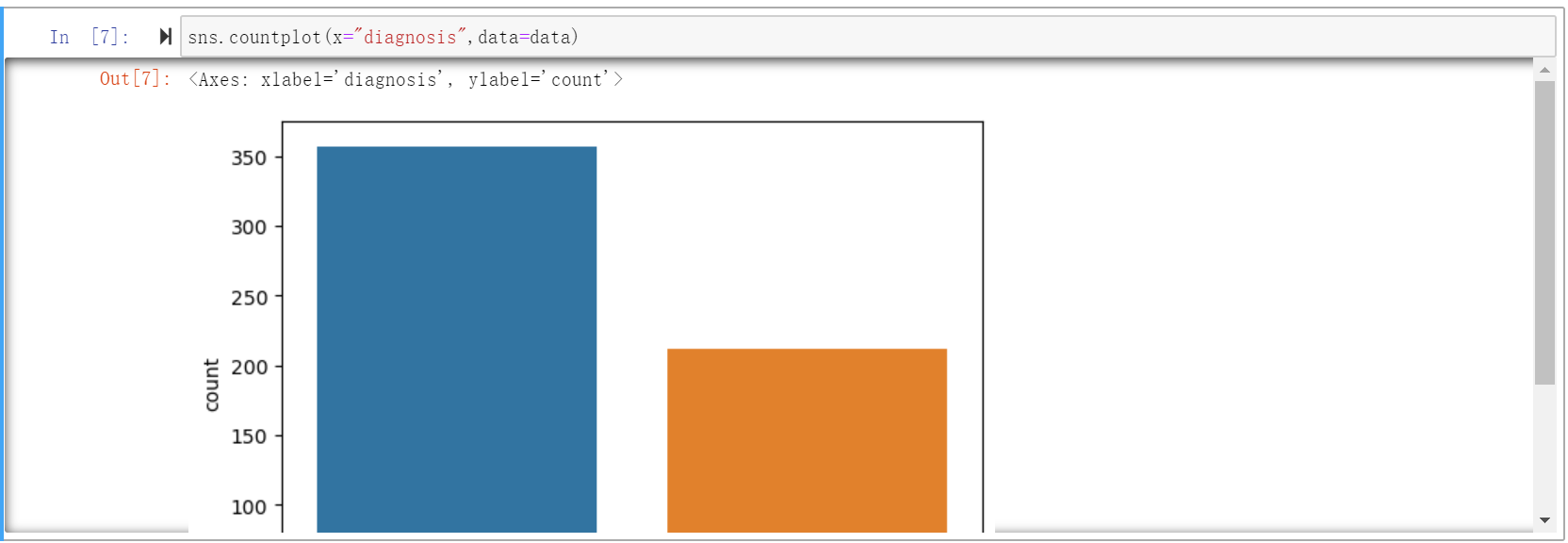
Convert the string variables 'M' and 'B' in the genus'diagnosis' to integer variables 1 and 0 using the map function.



Among them, 10 mean features, 10 standard deviation features and 10 maximum value features were extracted. Do machine learning with more important mean features.



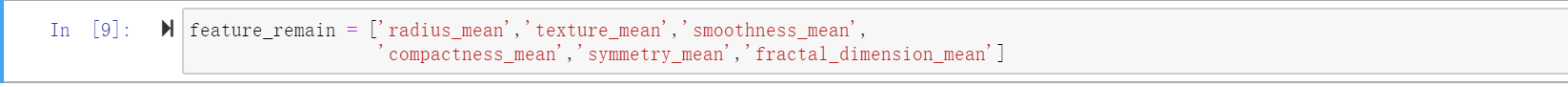
The sns.countplot() function was used to obtain the histograms of benign and malignant cells in the dataset.



The heat map sns.heatmap() function is used to obtain the correlation of each feature. The closer the value is to 1, the greater the correlation.

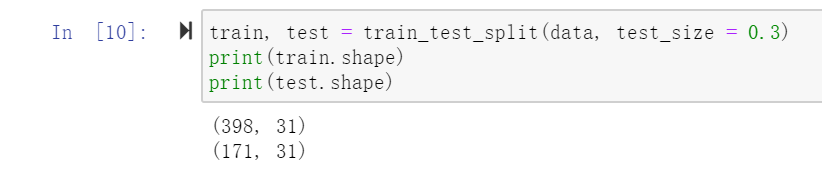


It can be seen from the heat map that the mean radius, mean perimeter and mean area have a high correlation, and compactness, concavity and concave point correlation are large, so the features can be condensed into 6.

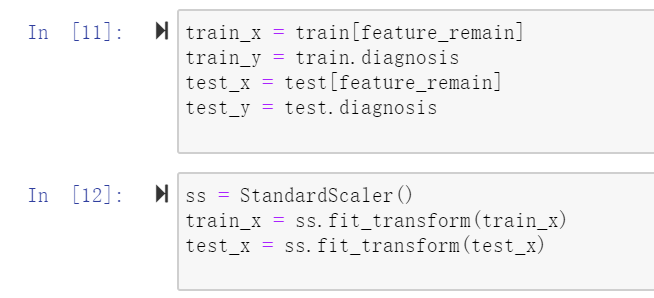


3.3.4 Support vector machine model building

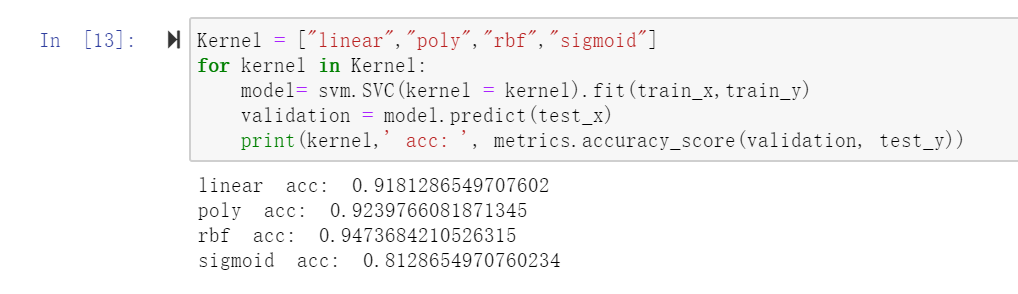
Divide the dataset into training and testing sets at a 7:3 ratio.



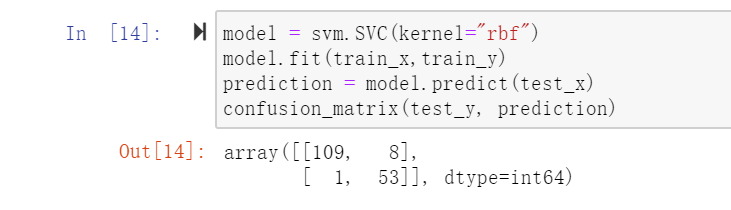
Define and normalize training and test datasets.



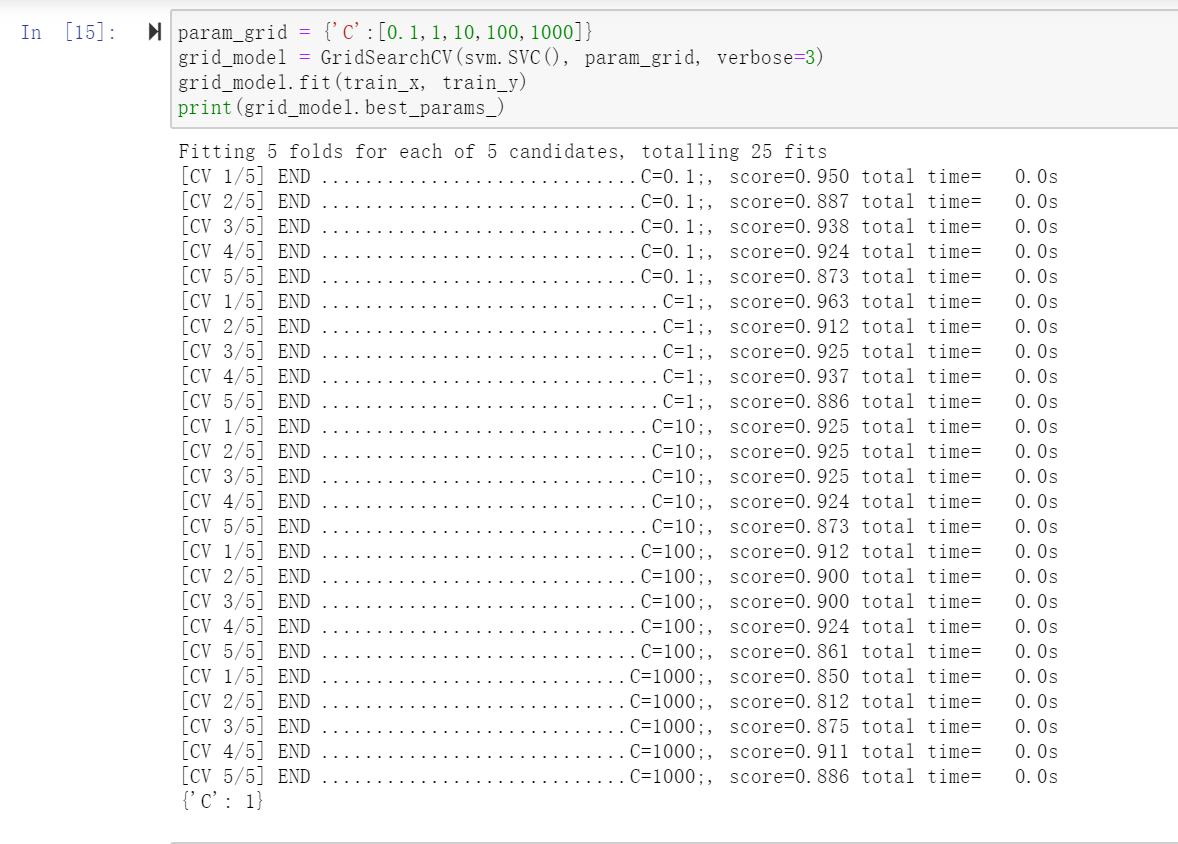
A support vector product model is established, and four kernel functions are used to train the data and test its accuracy. The Gaussian kernel function was found to have the highest accuracy of 94 736%。

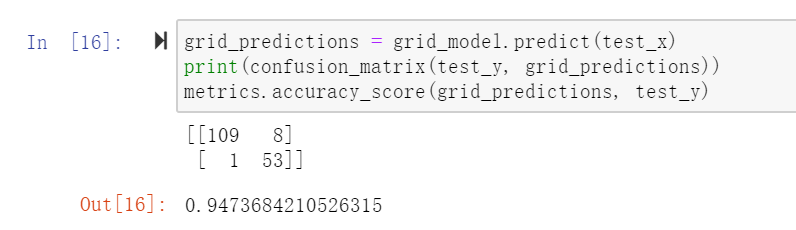


Lists the confusion matrix at this time.

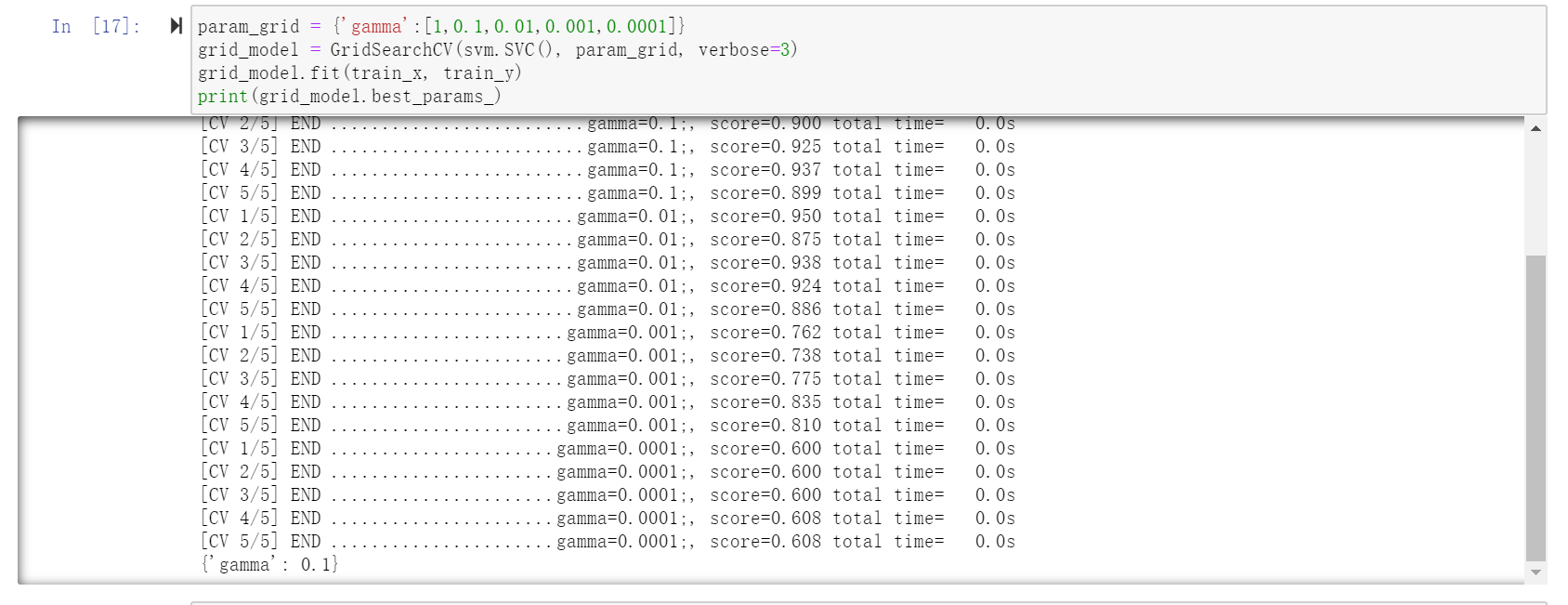


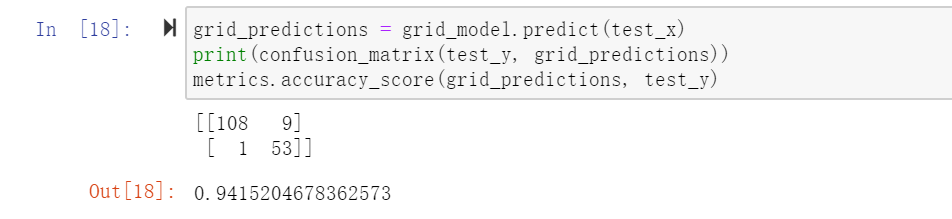
Adjust the penalty coefficient C in the support vector product model, and after adjustment, it is still the default value of 1, and the confusion matrix and accuracy rate remain unchanged.



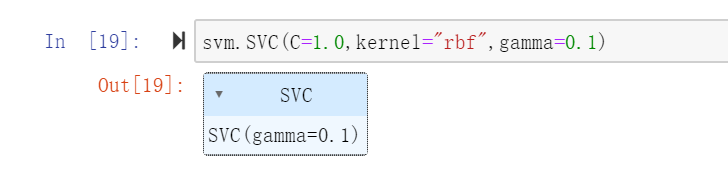


Adjust the parameter gamma in the support vector machine model, and take 0.1 after adjustment. The confusion matrix predicts the correct sample plus 1, and the accuracy rate is not improved.





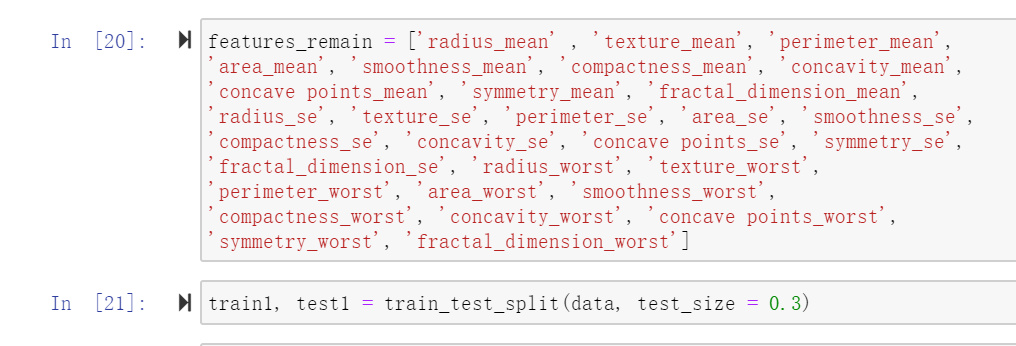
Get the support vector machine model model. The Gaussian kernel function is selected for the kernel function, 1 is taken for the penalty coefficient C, and 0.1 is taken for gamma.

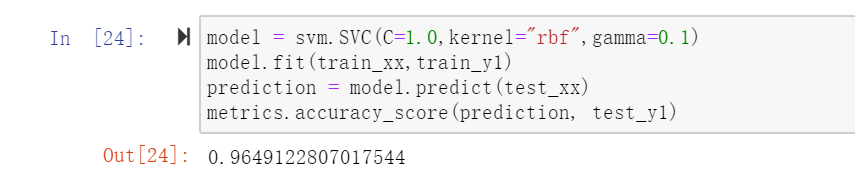


4 Experimental results and analysis

4.1 Experimental results

The dataset was trained and tested on all features of this model, and the accuracy rate was 96.491%.



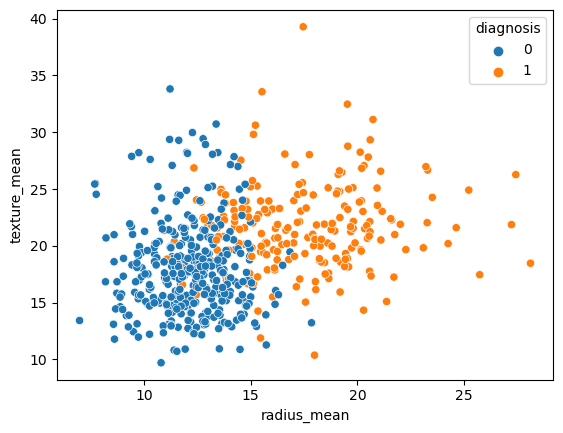


4.2 Data visualization and result analysis

Take the mean feature radius and texture mean as examples. The horizontal axis variable is the mean radius, the vertical axis variable is the texture mean, the blue point is benign, and the yellow point is malignant.

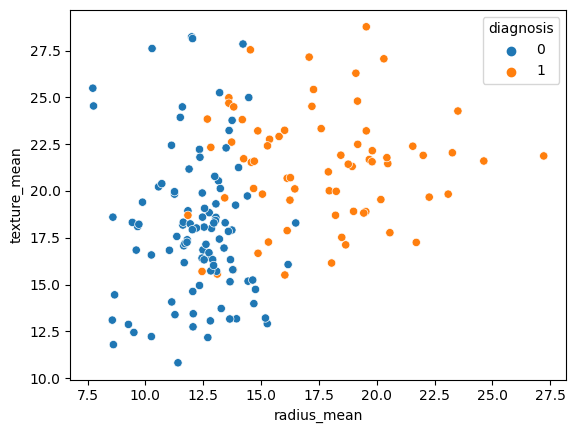
The feature distribution scatterplot of all data in the dataset is as follows:





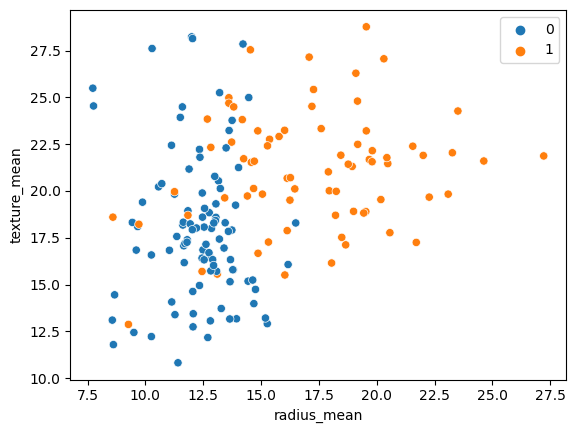
The feature distribution scatterplot of the test group is as follows:





The scatterplot of the feature distribution predicted by the model is as follows:





Obviously, the classification model is effective and has a high accuracy.

4.3 Summary

We use machine learning technology to establish a support vector machine model based on the characteristic data related to breast cancer molecular typing to predict and diagnose breast cancer, so that breast cancer can be detected and treated early. This has significant implications for the field of breast cancer disease medicine.