**基于SVM的乳腺癌数据集分类的设计与实现**

**摘要：**本篇报告研究了利用机器学习算法，通过乳腺肿块的检测数据自动进行诊断。数据集中共有569个样本，每个样本包含乳腺细胞的30个特征。这30个特征都是由数字化细胞核的10个基础特征（包括半径、质地、周长、面积、对称性等）的均值、标准差及最大值构成。本小组使用支持向量机，构建了一个进行乳腺癌诊断的分类器。经训练、测试，此模型准确率达96.491%。

**关键词**：乳腺癌，机器学习，分类模型，支持向量机

1 引言

乳腺癌具有高并发率和高死亡率的特点，严重威胁着全球女性的健康。

根据世界卫生组织（WHO）的报告，2018年全球共有209万乳腺癌病例，其中62.7万为死亡病例。乳腺癌已经成为全世界女性最常见的恶性肿瘤以及第二大癌症死因。

根据美国癌症协会的数据，1989年至2016年间，乳腺癌的死亡率已经降低了40%。2007年以来，50岁以下女性的死亡率并没有显著下降，但50岁以上女性的死亡率仍在继续下降。

这些进展要归功于更好的治疗方法以及乳腺癌的早期诊断。

目前，乳腺癌的分子分型需要通过穿刺活检来确定。本次实验利用机器学习技术，依据乳腺癌分子分型相关特征数据，建立一个分类器，以辅助进行乳腺癌预测及早期诊断。

2 问题的提出、方法与模型的介绍

本研究旨在于利用分类模型构建准确率尽可能高的分类器。

本数据集中类别有2个，分别为“良性”和“恶性”，细胞特征有30个。根据数据集的特点，我们选择了支持向量机为分类模型。

支持向量机（support vector machines，SVM）是一种二分类模型，SVM 的目的就是找一条线，以 “最好地” 区分这两类点，以至如果以后有了新的点，这条线也能做出很好的分类，这是在二维中的说明。在高维空间中，我们想要区分两类样本数据，需要找到一个超平面来区分两类样本数据。SVM 适合中小型数据样本、非线性、高维的分类问题。

“三八线”可以看作二维空间中SVM的形象解释，它传递出了以下几点重要的信息：

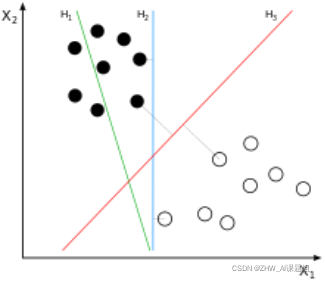
(1)是一条直线（线性函数）；

(2)能将桌面分为两个部分，分别属于你和我（具有分类功能，是一种二值分类）；

(3)位于课桌正中间，不偏向任何一方（注重公平原则，才能保证双方利益最大化）。

以上三点是SVM算法的中心思想。

SVM算法解释与分析：SVM 将会寻找可以区分两个类别并且能使间隔（margin）最大的划分超平面。比较好的划分超平面，样本局部扰动时对它的影响最小、产生的分类结果最鲁棒、对未见示例的泛化能力最强。由下图可知，H1是线性不可分的，H2、H3是线性可分的。这时，我们使用间隔最大的原理选择H3作为下图区分两类样本点的超平面。

****图四.svm超平面

事实上，大部分时候数据并不是线性可分的，这个时候满足这样条件的超平面就根本不存在。对于非线性的情况，SVM 的处理方法是选择一个核函数κ(.,.)，通过将数据映射到高维空间，最终在高维特征空间中构造出最优分离超平面，从而把平面上本身不好分的非线性数据分开。

3 实验原理及步骤

3.1 实验原理

记xi(i=1,2......30)为30个特征变量，[ai,yi](i=1,2......569)为样本，ai属于30维线性空间，良性yi=-1，恶性yi=-1。

定义最优划分(w\*X)+b=0,

X=[x1,x2,...x30],w属于30维空间，满足

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

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

其中，满足方程

(w\*ai)+b=+-1的样本为支持向量。

要使两类总体的分类面距离最大，则有

Max 2/|w|

即

Min |w|^2/2

于是建立SVM的如下数学模型：

Min |w|^2/2

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

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

求得最优值对应的w和b,可得分类函数

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

3.2实验步骤

3.2.1数据集

本研究实验数据集来自威斯康星州诊断性乳腺癌 (WDBC)数据集。这些特征是从细针抽吸（FNA）乳房肿块的数字化图像中计算出来的。它们描述了图像中存在的细胞核特征。

实例数：569

属性数：32（ID、诊断、30个实值特征）

属性信息

1) 身份证号码

2) 诊断（M = 恶性，B = 良性）

实值特征：

a) 半径

b) 纹理

c) 周界

d) 面积

e) 平滑度

f) 紧凑度

g) 凹度

h) 凹点

i) 对称性

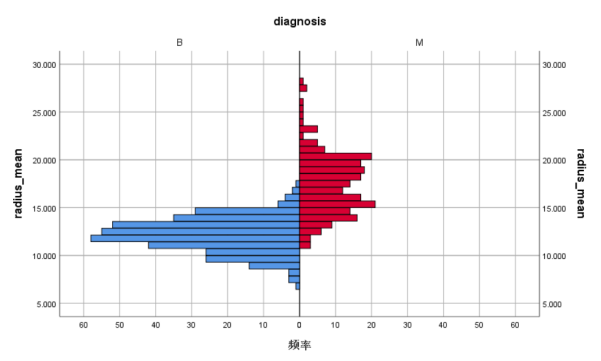
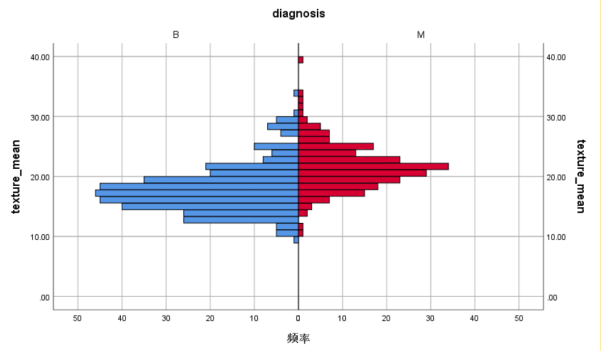
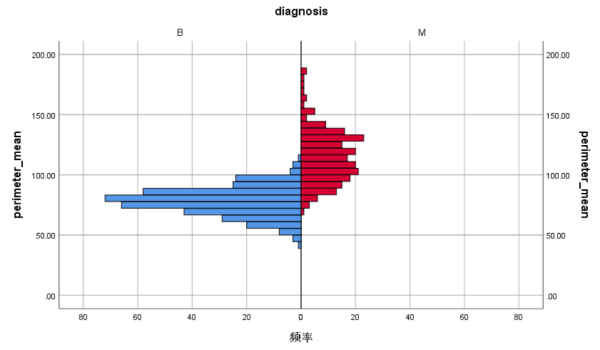
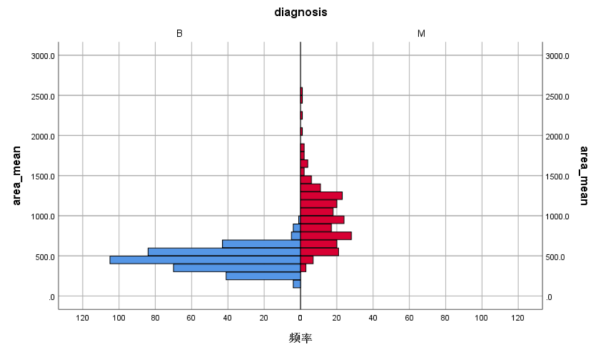
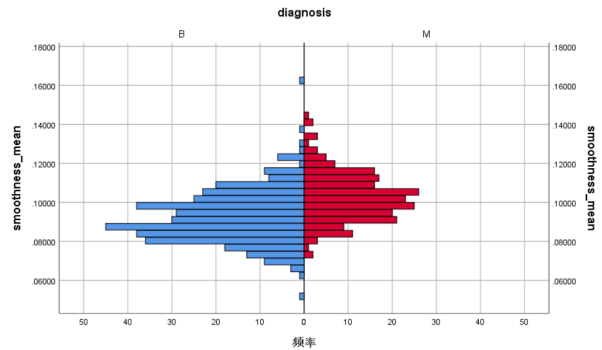
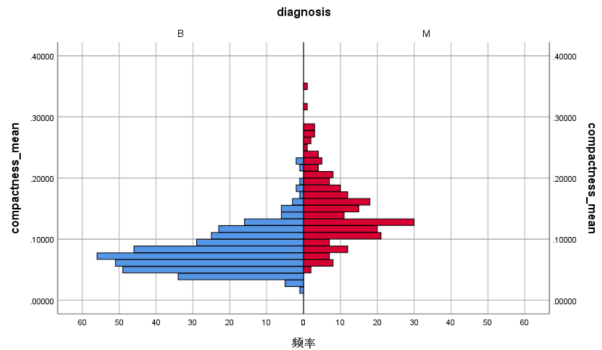
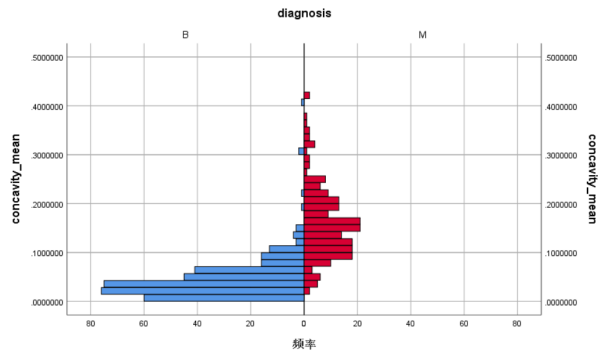
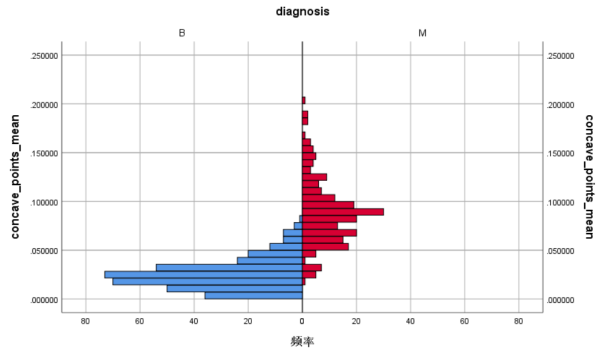
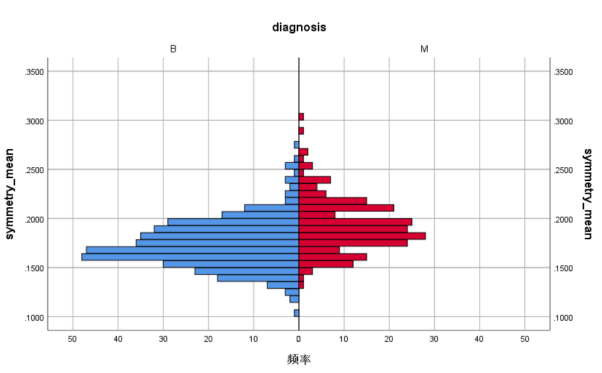
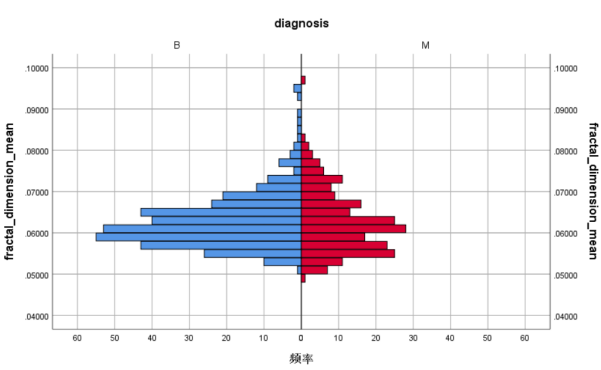
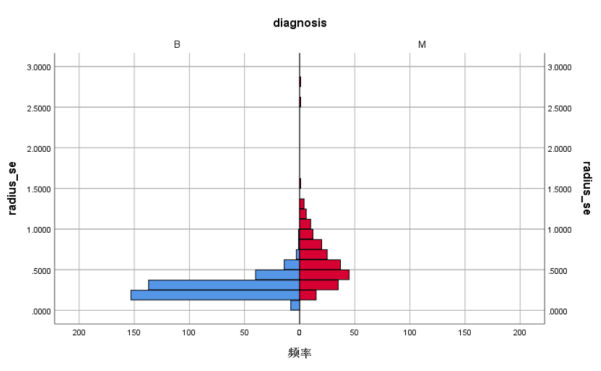
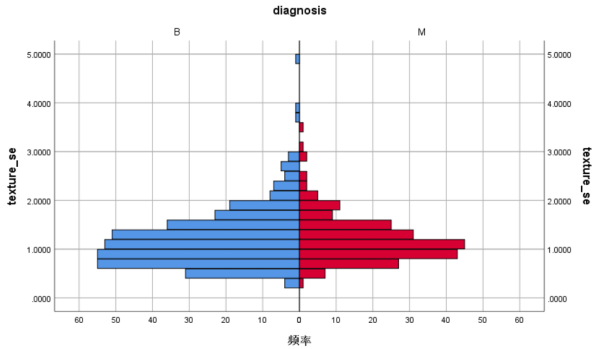
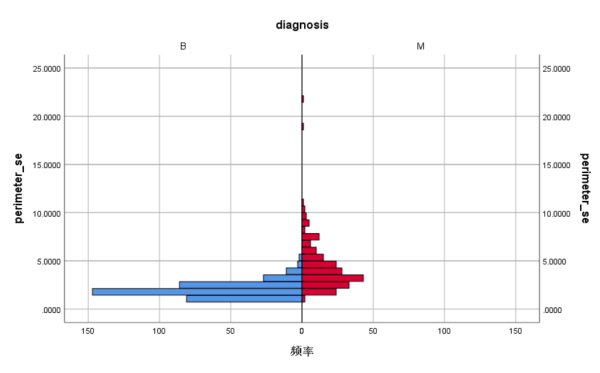
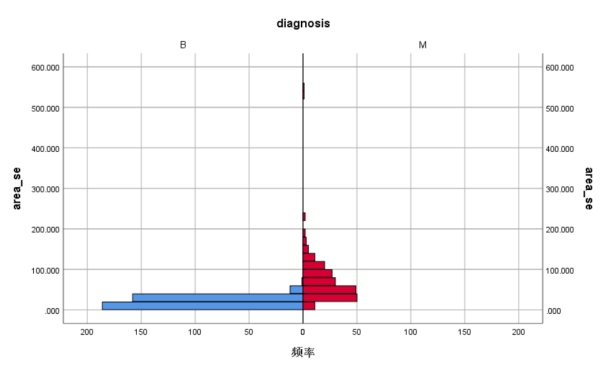
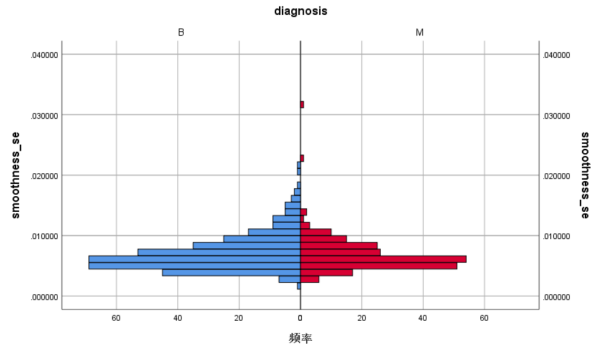
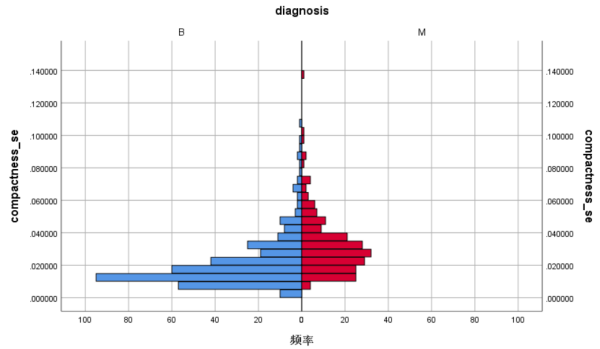
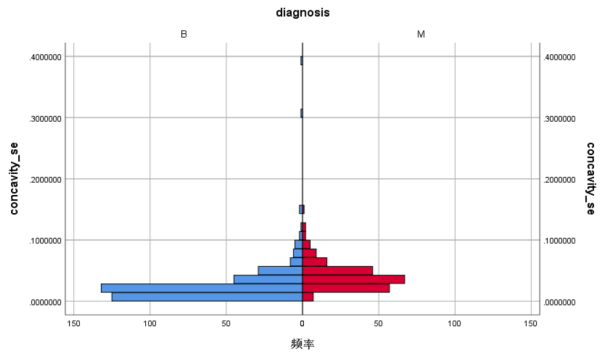
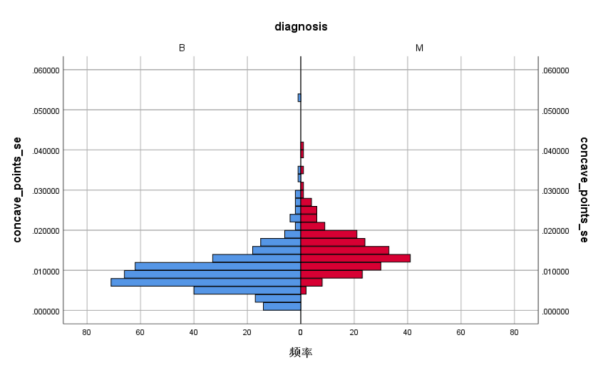
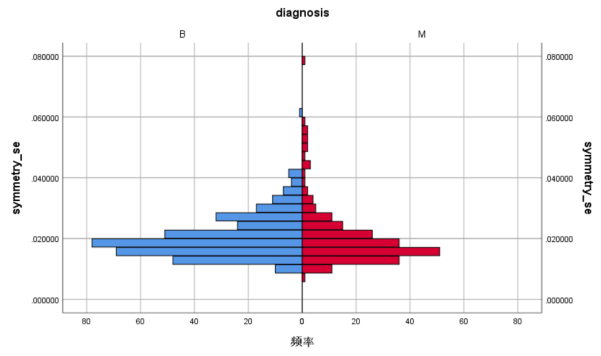
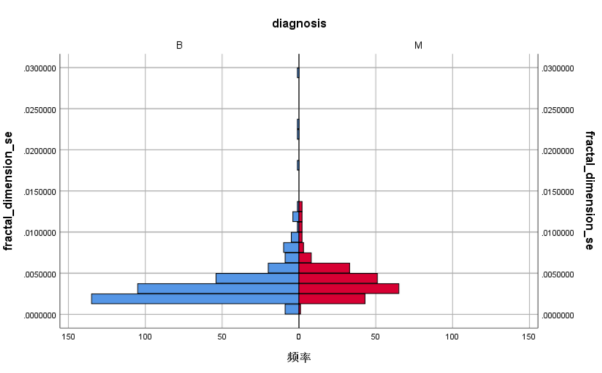
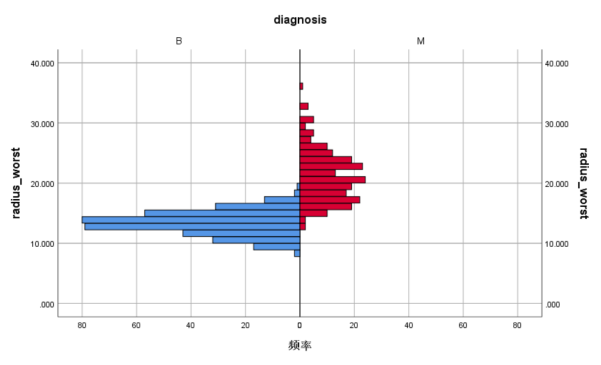
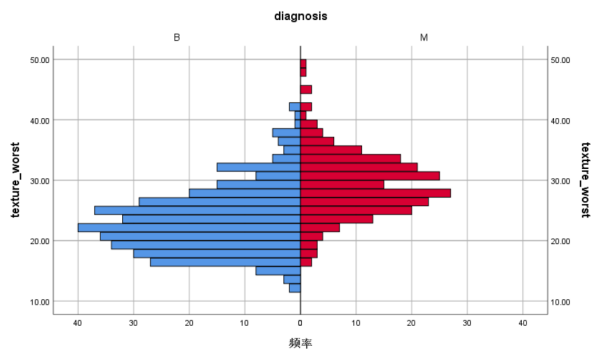
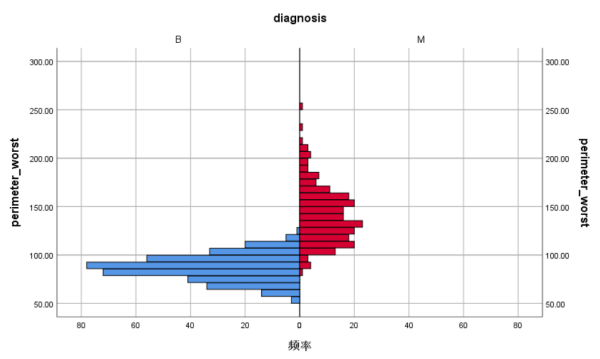
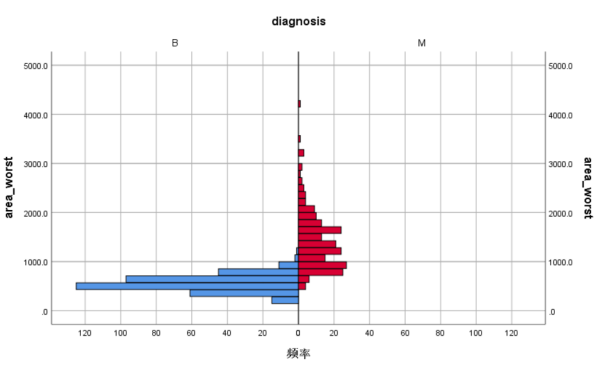
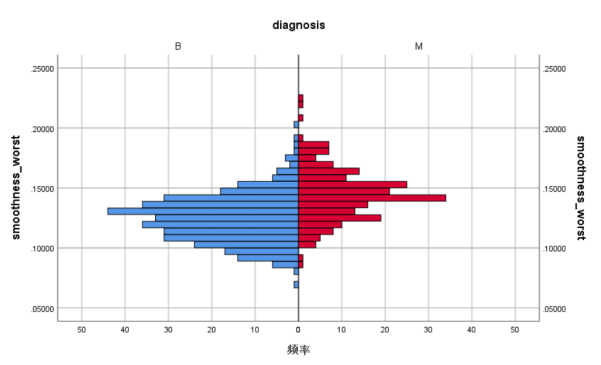
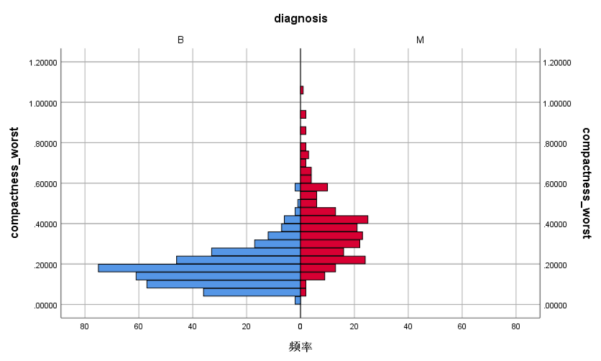
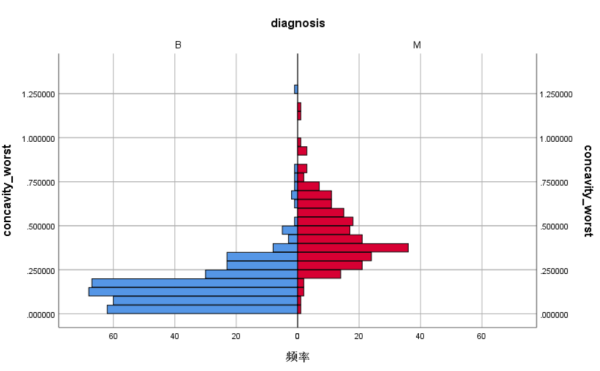
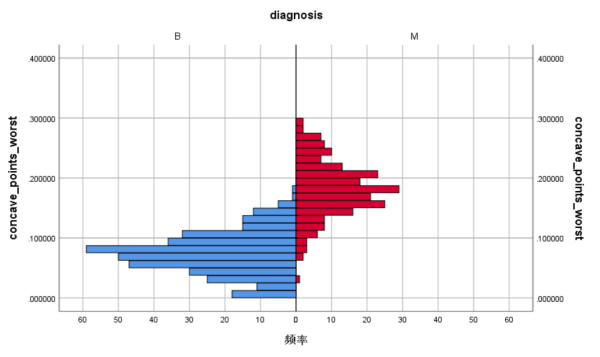
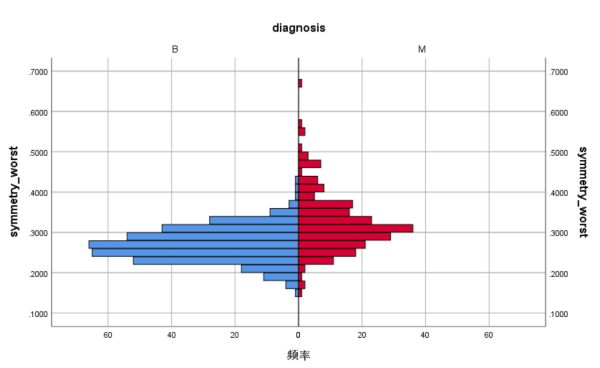
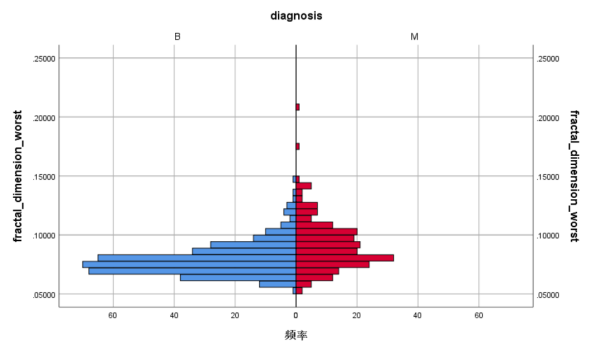
j) 分形维数

以上实值特征的均值、标准差和最大值（最差）

我们把提供的data文件转为csv文件并将特征名称加入第一行。

3.2.2数据预处理

首先通过“人口金字塔”图分析数据集中的各项特征。

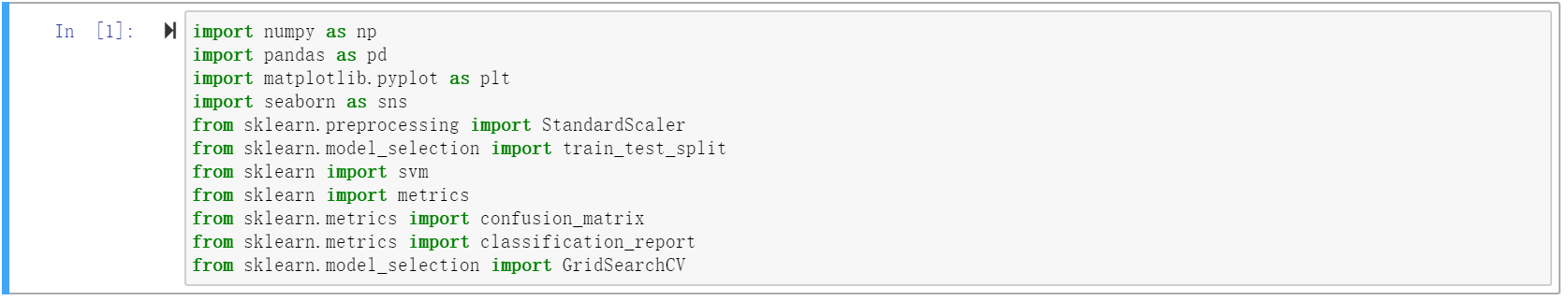
1. 半径均值
2. 纹理均值
3. 周界均值
4. 面积均值
5. 平滑度均值
6. 紧凑度均值
7. 凹度均值
8. 凹点均值
9. 对称性均值
10. 分形维数均值
11. 半径标准差
12. 纹理标准差
13. 周界标准差
14. 面积标准差
15. 平滑度标准差
16. 紧凑度标准差
17. 凹度标准差
18. 凹点标准差
19. 对称性标准差
20. 分形维数标准差
21. 半径最大值
22. 纹理最大值
23. 周界最大值
24. 面积最大值
25. 平滑度最大值
26. 紧凑度最大值
27. 凹度最大值
28. 凹点最大值
29. 对称性最大值
30. 分形维数最大值

通过以上30个特征的类间区分度分析。不难发现半径、周界、面积和凹点、凹度区分度较明显，而各属性的均值区分度也较高。特征重要性排序如下：

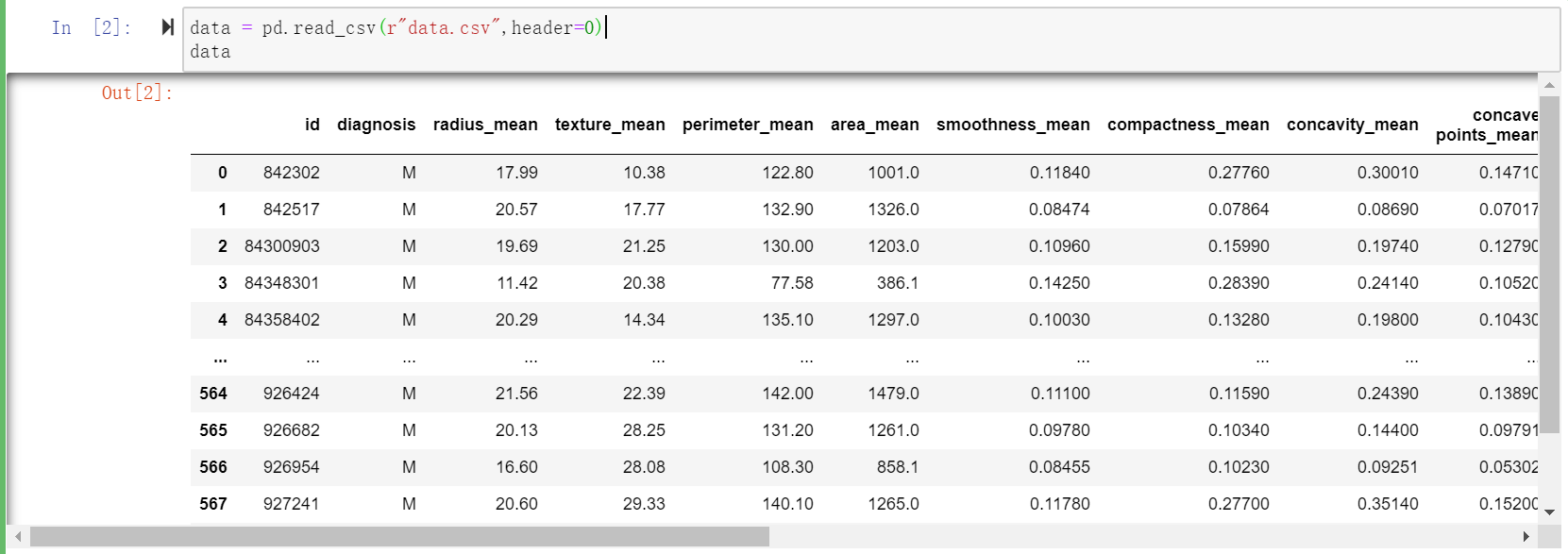
半径均值=周界均值=面积均值>凹点均值=凹度均值

3.3.3预处理代码

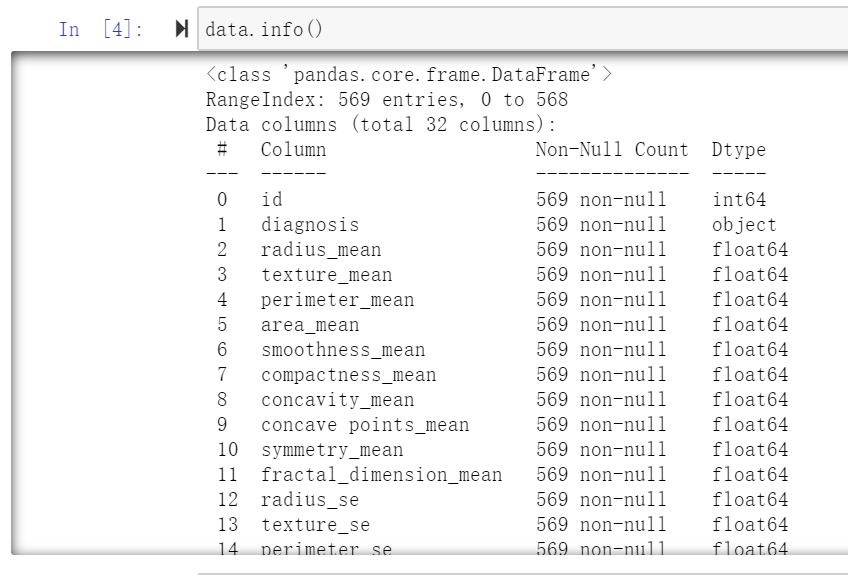
导入要用到的库



打开数据集



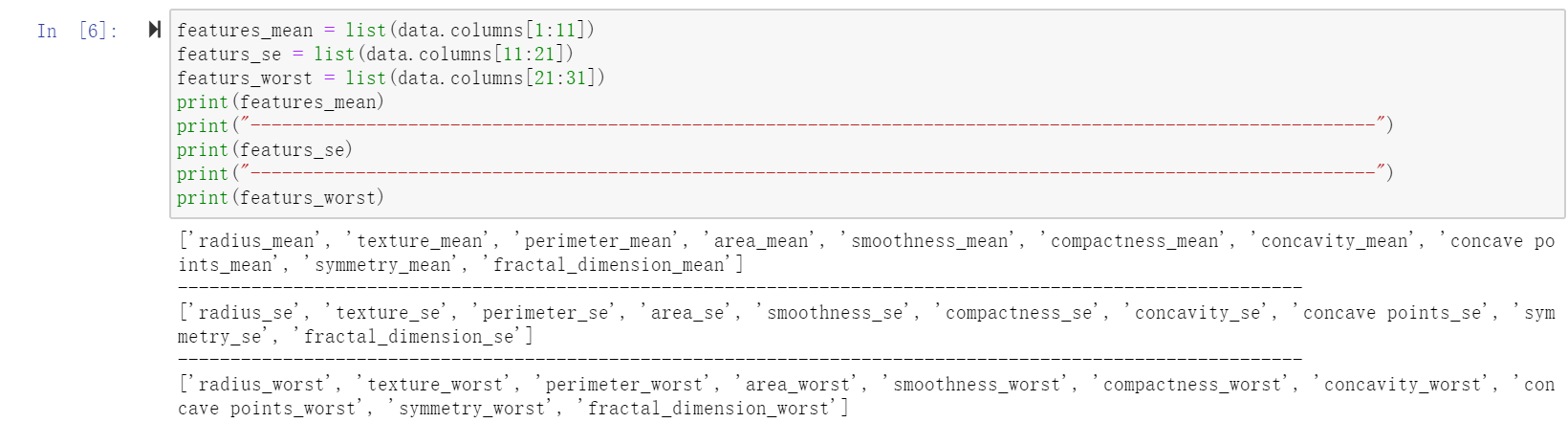
检查数据集中的特征值及数据类型



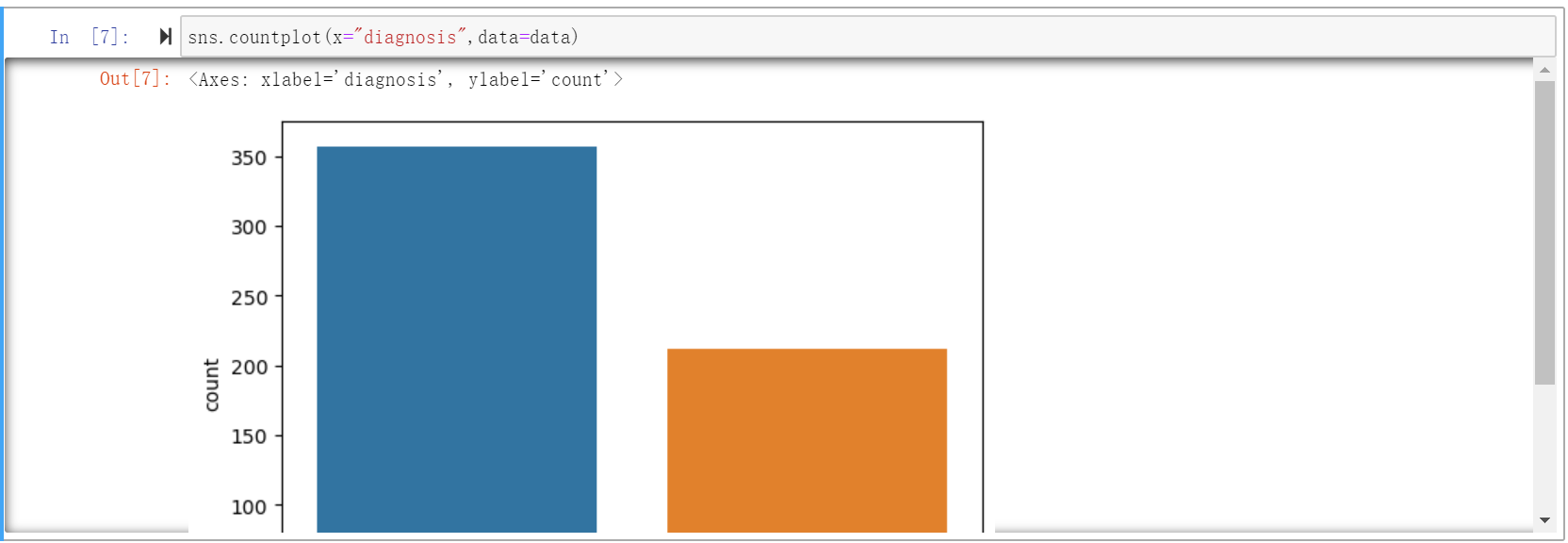
将属’diagnosis’中字符串变量’M’和’B’用map函数转化为整数变量1和0。



提取其中10个均值特征，10个标准差特征和10个最大值特征。用重要性更强的均值特征做机器学习。



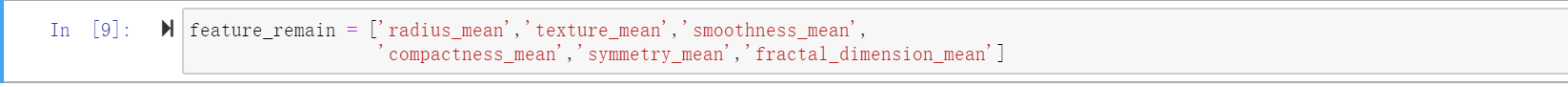
利用sns.countplot()函数得到数据集中细胞良性与恶性的直方图。



利用热力图sns.heatmap()函数得到各特征的相关性。数值越接近1相关性越大。

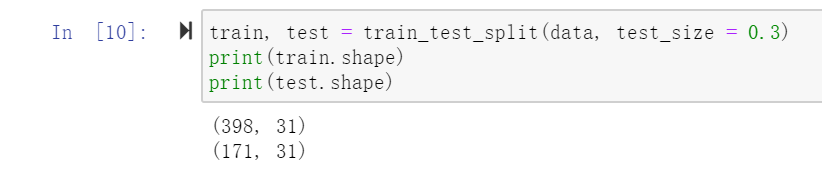


由热力图可知，半径均值、周界均值和面积均值相关程度高，紧凑度、凹度和凹点相关程度大，于是特征可以被浓缩为6个。

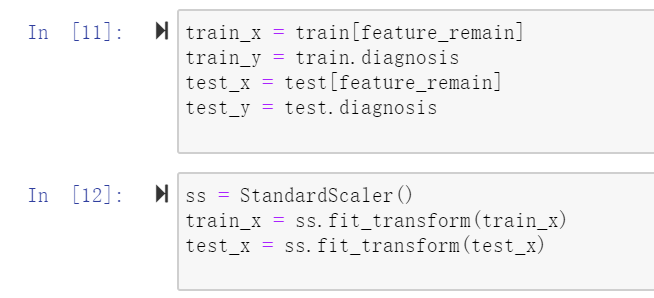


3.3.4支持向量机模型建立

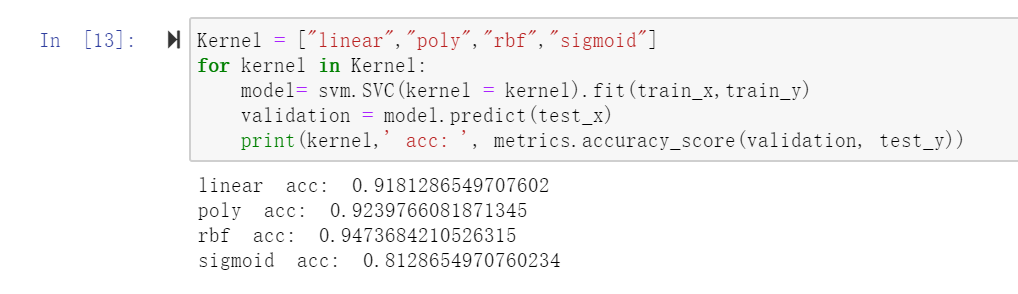
以7：3比例划分数据集为训练集和测试集。



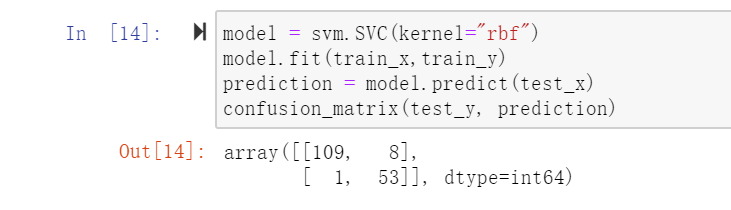
定义训练及测试数据集，并将它们归一化。



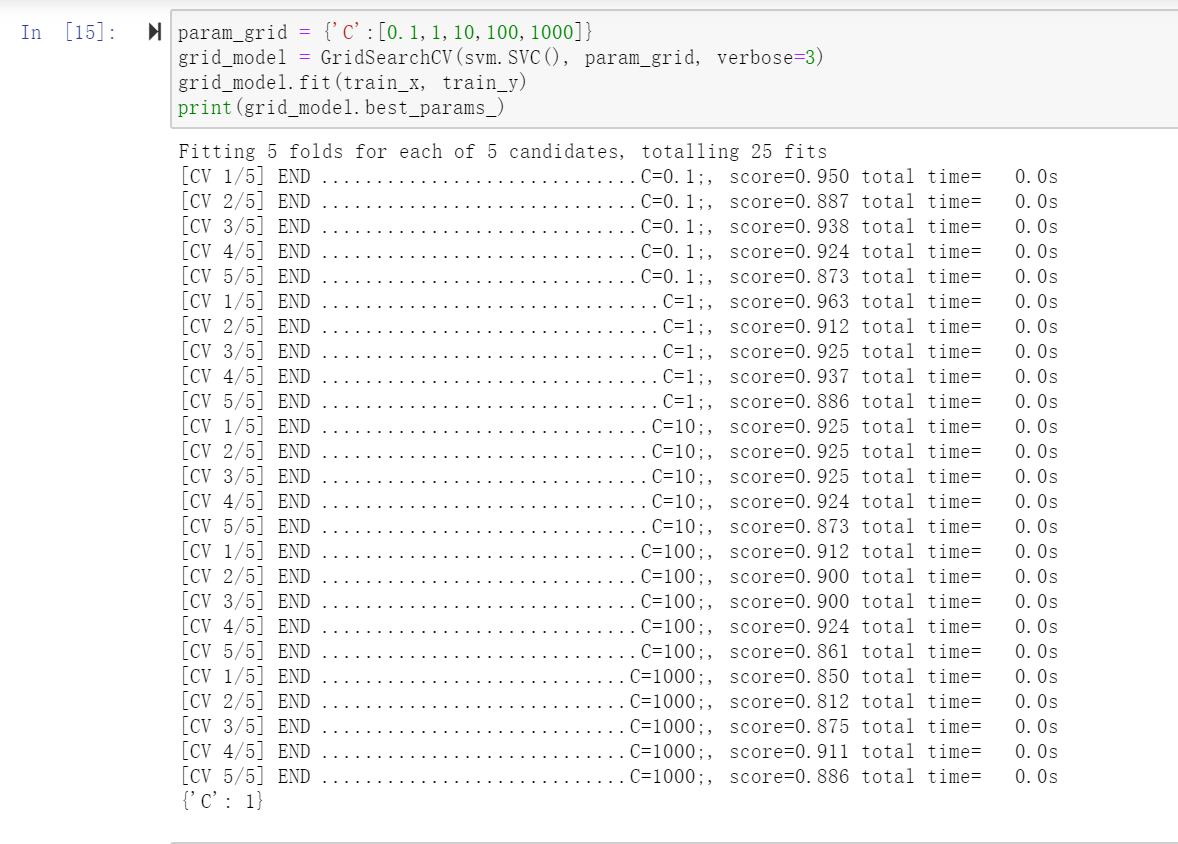
建立支持向量积模型，利用四种核函数对数据进行训练并测试其准确率。发现高斯核函数准确率最高，为94.736%。

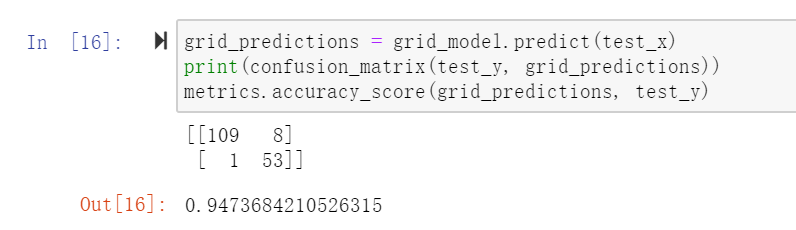


列出此时的混淆矩阵。

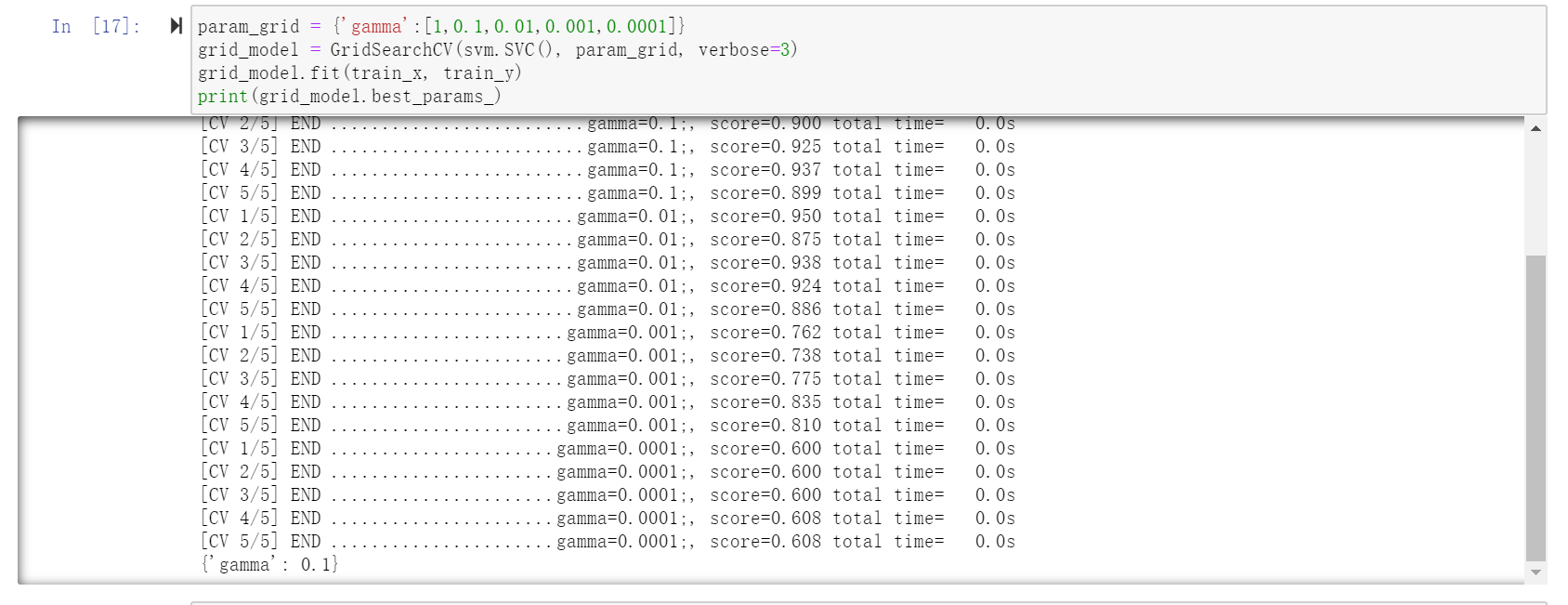


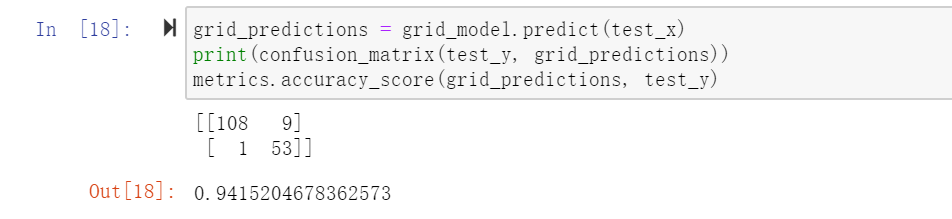
调整支持向量积模型中的罚系数C，调整后仍为默认值1，混淆矩阵与准确率不变。



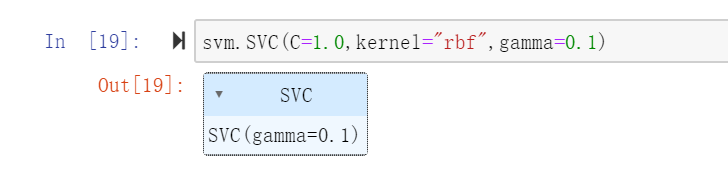


调整支持向量机模型中的参数gamma，调整后gamma取0.1。混淆矩阵预测正确样本加1，准确率未提高。





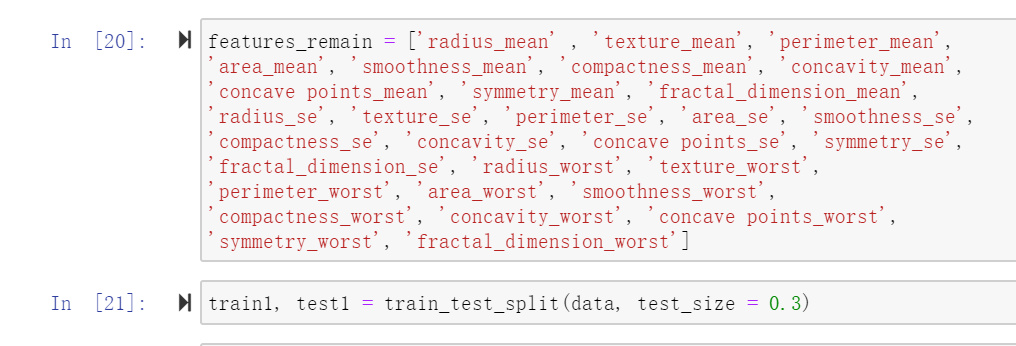
得到支持向量机模型model。其中核函数选择高斯核函数，罚系数C取1，gamma取0.1。

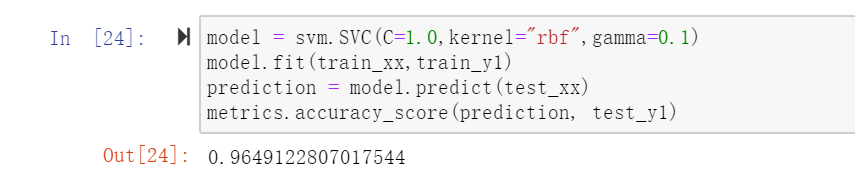


4实验结果及分析

4.1实验结果

将数据集对所有特征以此模型进行训练和测试，得到准确率为96.491%。



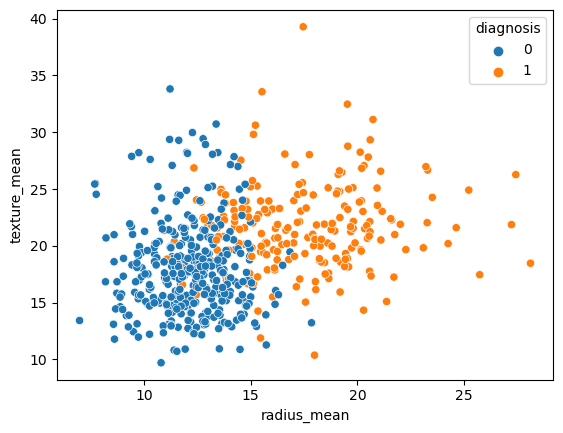


4.2数据可视化及结果分析

以特征半径均值和纹理均值为例。横轴变量为半径均值，纵轴变量为纹理均值，蓝色点为良性，黄色点为恶性。

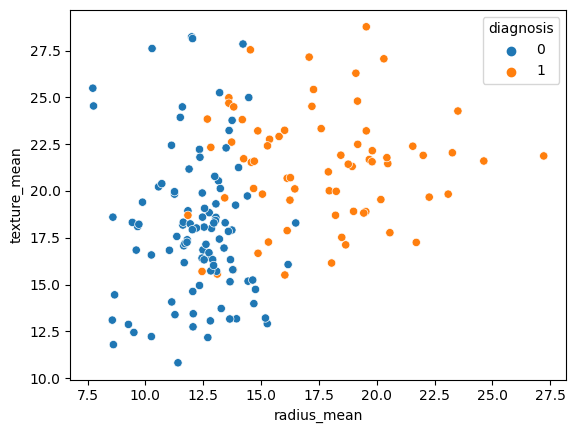
数据集中所有数据的特征分布散点图如下：





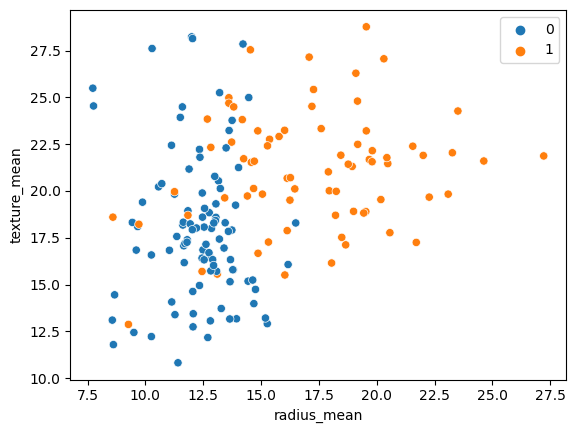
测试组的特征分布散点图如下：





根据该模型所预测的特征分布散点图如下：





显然，该分类模型有效，且准确率较高。

4.3总结

我们利用机器学习技术，依据乳腺癌分子分型相关特征数据，建立一个支持向量机模型，进行乳腺癌预测及早期诊断，使得乳腺癌能够早发现早治疗。这对乳腺癌疾病医学领域意义非凡。