构建逻辑回归模型对乳腺癌进行诊断，描述主要过程并粘贴实现代码。

数据集下载地址为：https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

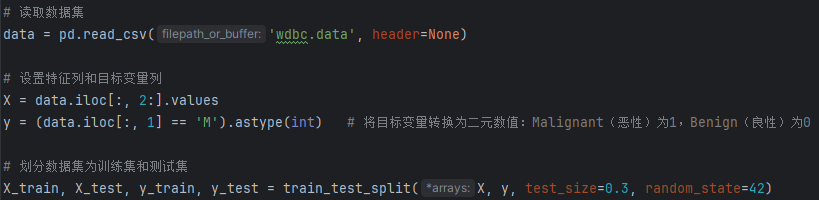
**构建逻辑回归模型对乳腺癌进行诊断**

**1.数据预处理**

特征和目标变量的设定：选择数据集中第3列到最后一列作为特征（X），第2列作为目标变量（y），并将目标变量中的'M'标记转换为1，代表恶性肿瘤，'B'标记转换为0，代表良性肿瘤。

数据集划分：使用train\_test\_split函数将数据集划分为训练集和测试集，测试集占30%，随机状态设置为42以保证结果的可复现性。

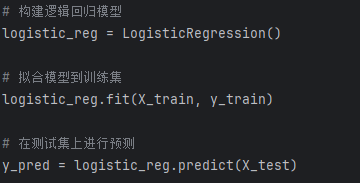
代码：



**2.模型构建与训练**

逻辑回归模型初始化：创建一个逻辑回归模型实例，使用LogisticRegression构建逻辑回归模型。

模型拟合：使用训练集数据拟合（训练）逻辑回归模型，在测试集上进行预测。计算模型的准确率、混淆矩阵和分类报告。

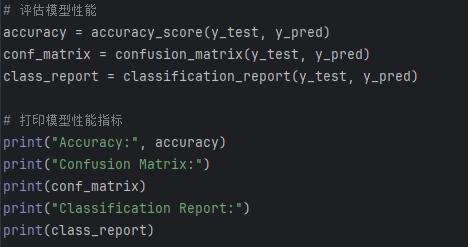
代码：  


**3.模型评估**

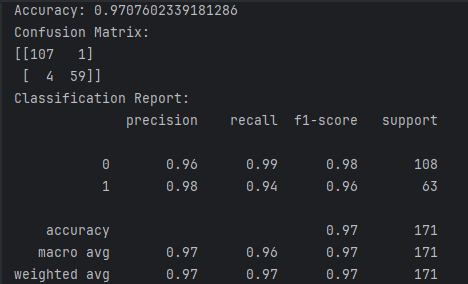
预测：使用训练好的模型对测试集进行预测。

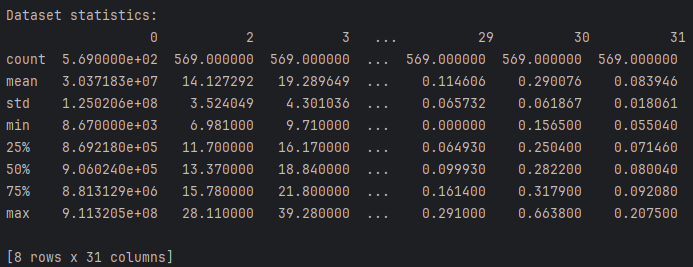
性能评估：计算模型的准确率、混淆矩阵和分类报告，以评估模型的性能。准确率表示正确预测的样本占总样本的比例；混淆矩阵展示了模型预测的真正例（TP）、预测的假正例（FP）、预测的真反例（TN）和预测的假反例（FN）；分类报告提供了模型的精确率、召回率、F1分数等详细性能指标。

代码：



代码输出结果：





**4.可视化特征重要性**

特征系数获取：从逻辑回归模型中获取特征系数，这些系数表示特征对模型预测的贡献大小。计算逻辑回归模型的特征系数。

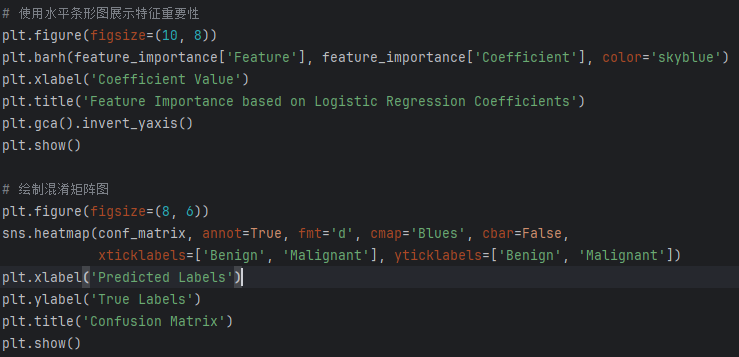
特征重要性可视化：创建一个DataFrame来存储特征名称和对应的系数，然后根据系数的绝对值对特征进行排序，并使用水平条形图可视化特征的重要性。

代码：  

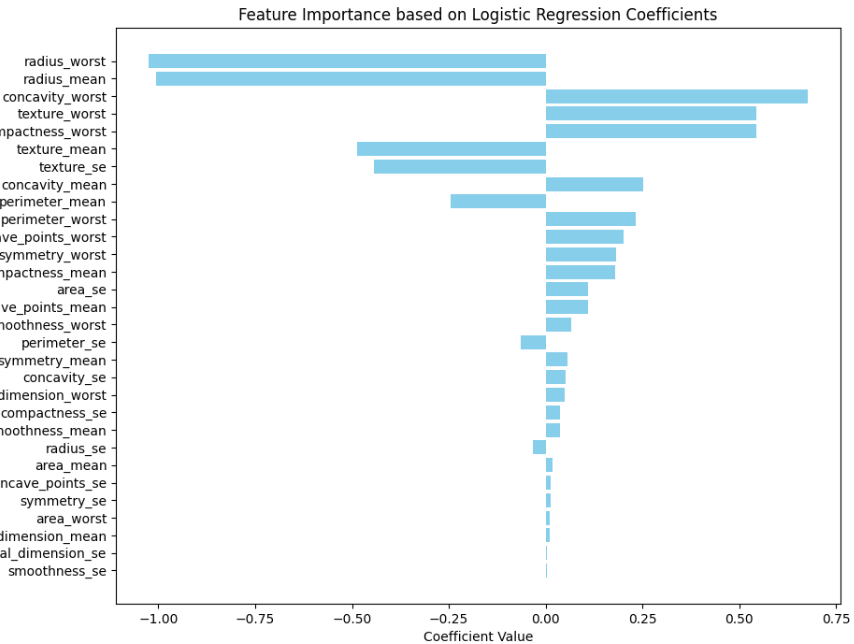

**5.混淆矩阵可视化：**

绘制混淆矩阵，显示模型的预测结果与真实标签之间的对应关系。

代码：



代码输出结果：



**6.数据集统计和特征相关性热图：**

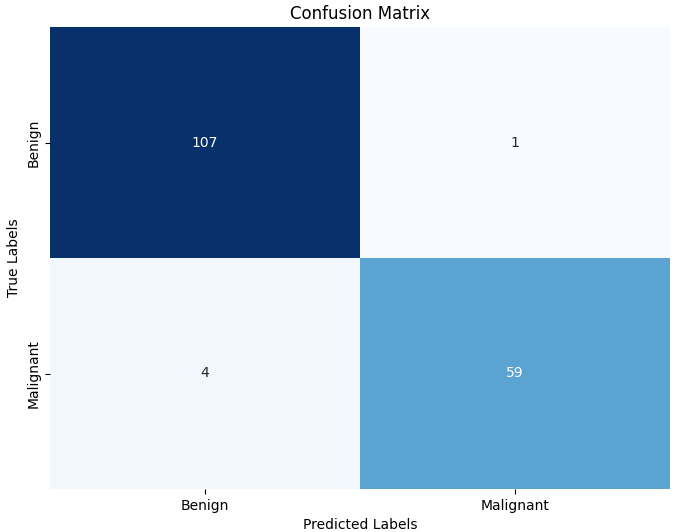
数据集统计摘要：打印数据集的统计摘要，包括均值、标准差、最小值、四分位数和最大值等。

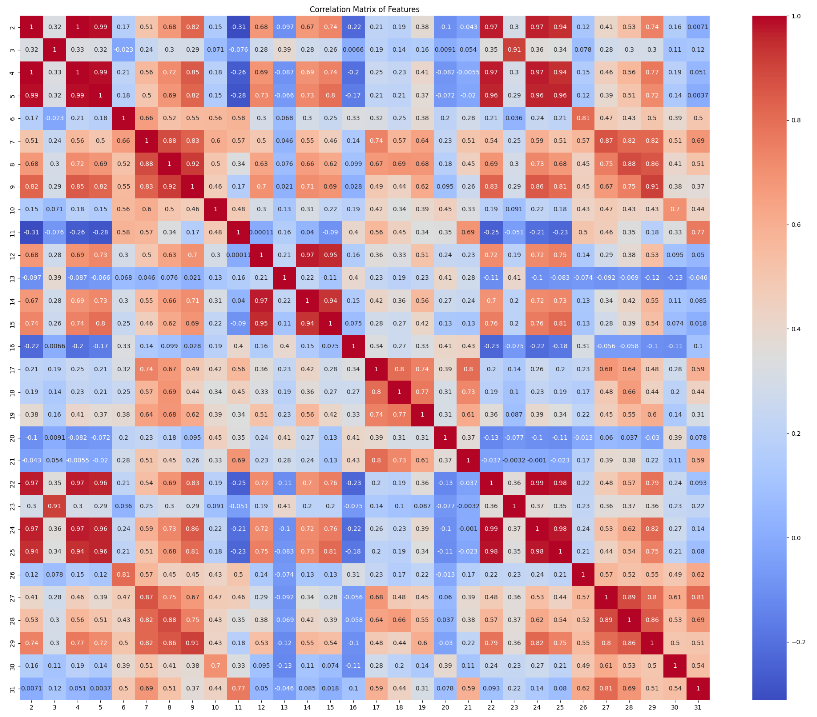
相关性分析：计算特征之间的相关性矩阵，并使用热图可视化，以识别特征间的线性关系，展示特征之间的相关性。

代码：



代码输出结果：





**结果分析**

准确率（Accuracy）：模型的准确率约为97%，这表明模型在绝大多数情况下能够做出正确的预测，显示出在乳腺癌诊断方面的高效率。

混淆矩阵（Confusion Matrix）：混淆矩阵显示，模型预测出107个真阴性（TN）和59个真阳性（TP），同时存在1个假阴性（FN）和4个假阳性（FP）。这意味着模型在识别良性肿瘤方面表现尤为出色，但在恶性样本的识别上存在极少量误判。

分类报告（Classification Report）：分类报告提供了包括精确率、召回率和F1分数在内的详细指标。模型在良性和恶性样本上的精确率和召回率都相当高，但需要注意到恶性样本的数量较少，可能导致模型在这一子集上的表现不够稳定。

具体数值如下：

真阴性 (True Negative, TN): 107

假阴性 (False Negative, FN): 1

假阳性 (False Positive, FP): 4

真阳性 (True Positive, TP): 59

混淆矩阵的分析表明，尽管模型在恶性肿瘤的识别上存在极个别的漏诊和误诊，但整体表现非常出色。

特征重要性分析：

texture\_worst：细胞核最差纹理，高纹理值可能指示癌细胞的异质性，是恶性肿瘤的潜在标志。

radius\_se：细胞核半径的标准误，较大值表明细胞核尺寸变异性大，通常与癌症发展相关。

symmetry\_worst：细胞核最差对称性，癌细胞的不对称性可能与更高的侵袭性相关。

concave\_points\_mean：细胞核平均凹点数量，较多的凹点可能与恶性肿瘤相关。

concavity\_worst：细胞核最差凹度，较深的凹陷可能指示更具侵袭性的癌细胞。

绝对值大的数值代表，哪些特性对诊断有较大的影响。

此外，通过热图分析特征间的相关性，我们发现：颜色的深浅表示相关性的强度，其中暖色表示强正相关，冷色表示强负相关。通过热图，我们可以识别出一些特征间的强相关性，这可能提示我们在特征选择时考虑去除冗余特征，以提高模型的泛化能力。

**结论**

逻辑回归模型在乳腺癌诊断任务上表现出色，具有高准确率和良好的分类性能。特征重要性分析揭示了在诊断中起关键作用的特征，这些特征的医学意义为医疗专业人员提供了有价值的信息，有助于更准确地理解和诊断乳腺癌。尽管模型在恶性样本的识别上还有改进空间，但整体而言，模型提供了一个强大的工具，以辅助乳腺癌的早期诊断和治疗决策。

源代码：

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

import seaborn as sns

# 读取数据集

data = pd.read\_csv('wdbc.data', header=None)

# 设置特征列和目标变量列

X = data.iloc[:, 2:].values

y = (data.iloc[:, 1] == 'M').astype(int) # 将目标变量转换为二元数值：Malignant（恶性）为1，Benign（良性）为0

# 划分数据集为训练集和测试集

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# 构建逻辑回归模型

logistic\_reg = LogisticRegression()

# 拟合模型到训练集

logistic\_reg.fit(X\_train, y\_train)

# 在测试集上进行预测

y\_pred = logistic\_reg.predict(X\_test)

# 评估模型性能

accuracy = accuracy\_score(y\_test, y\_pred)

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

class\_report = classification\_report(y\_test, y\_pred)

# 打印模型性能指标

print("Accuracy:", accuracy)

print("Confusion Matrix:")

print(conf\_matrix)

print("Classification Report:")

print(class\_report)

# 获取逻辑回归模型的特征系数

coefficients = logistic\_reg.coef\_[0]

# 特征名称

feature\_names = [

"radius\_mean", "texture\_mean", "perimeter\_mean", "area\_mean", "smoothness\_mean",

"compactness\_mean", "concavity\_mean", "concave\_points\_mean", "symmetry\_mean", "fractal\_dimension\_mean",

"radius\_se", "texture\_se", "perimeter\_se", "area\_se", "smoothness\_se",

"compactness\_se", "concavity\_se", "concave\_points\_se", "symmetry\_se", "fractal\_dimension\_se",

"radius\_worst", "texture\_worst", "perimeter\_worst", "area\_worst", "smoothness\_worst",

"compactness\_worst", "concavity\_worst", "concave\_points\_worst", "symmetry\_worst", "fractal\_dimension\_worst"

]

# 创建一个DataFrame来存储特征名称和对应的系数

feature\_importance = pd.DataFrame({

'Feature': feature\_names,

'Coefficient': coefficients

})

# 根据系数的绝对值对特征进行排序

feature\_importance = feature\_importance.reindex(feature\_importance.Coefficient.abs().sort\_values(ascending=False).index)

# 使用水平条形图展示特征重要性

plt.figure(figsize=(10, 8))

plt.barh(feature\_importance['Feature'], feature\_importance['Coefficient'], color='skyblue')

plt.xlabel('Coefficient Value')

plt.title('Feature Importance based on Logistic Regression Coefficients')

plt.gca().invert\_yaxis()

plt.show()

# 绘制混淆矩阵图

plt.figure(figsize=(8, 6))

sns.heatmap(conf\_matrix, annot=True, fmt='d', cmap='Blues', cbar=False,

xticklabels=['Benign', 'Malignant'], yticklabels=['Benign', 'Malignant'])

plt.xlabel('Predicted Labels')

plt.ylabel('True Labels')

plt.title('Confusion Matrix')

plt.show()

# 显示数据集统计信息

print("Dataset statistics:\n", data.describe())

# 热力相关矩阵

corr = data.iloc[:, 2:].astype(float).corr()

plt.figure(figsize=(25, 20))

sns.heatmap(corr, annot=True, cmap='coolwarm')

plt.title('Correlation Matrix of Features')

plt.show()