**实验一**

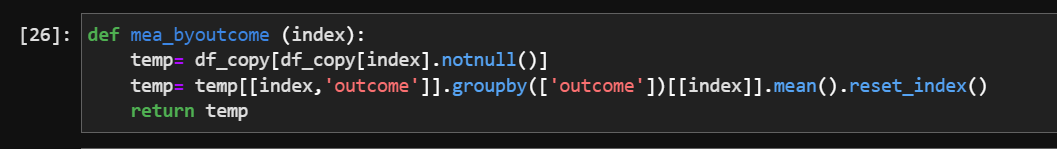
**1.机器学习：回归问题**

（1）数据集的特征变量里glucose、blood\_pressure、skin\_thick、insulin、BMI这五项的最小数据不应为零，应将其转换为NaN值，同时计算出各列的平均值并将其填入缺失值。

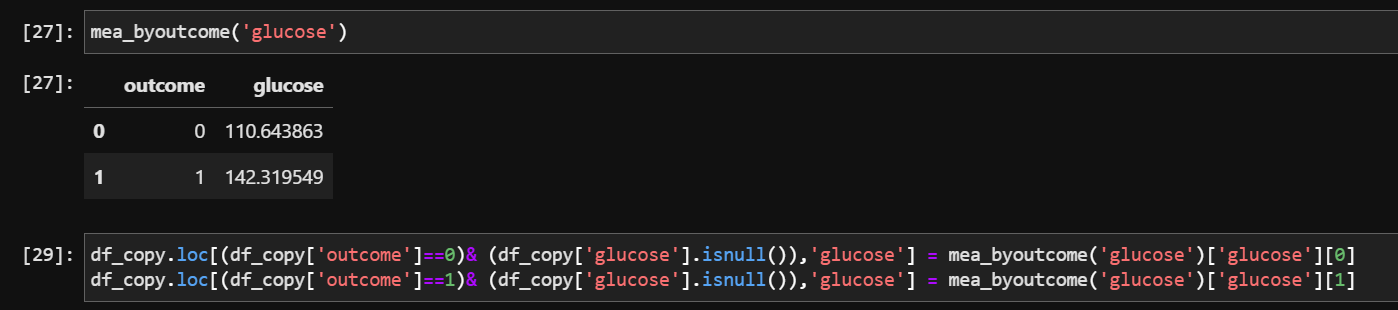
零值转换为NaN值



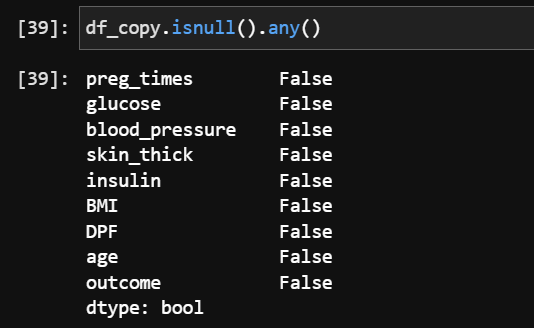
计算各列均值的函数



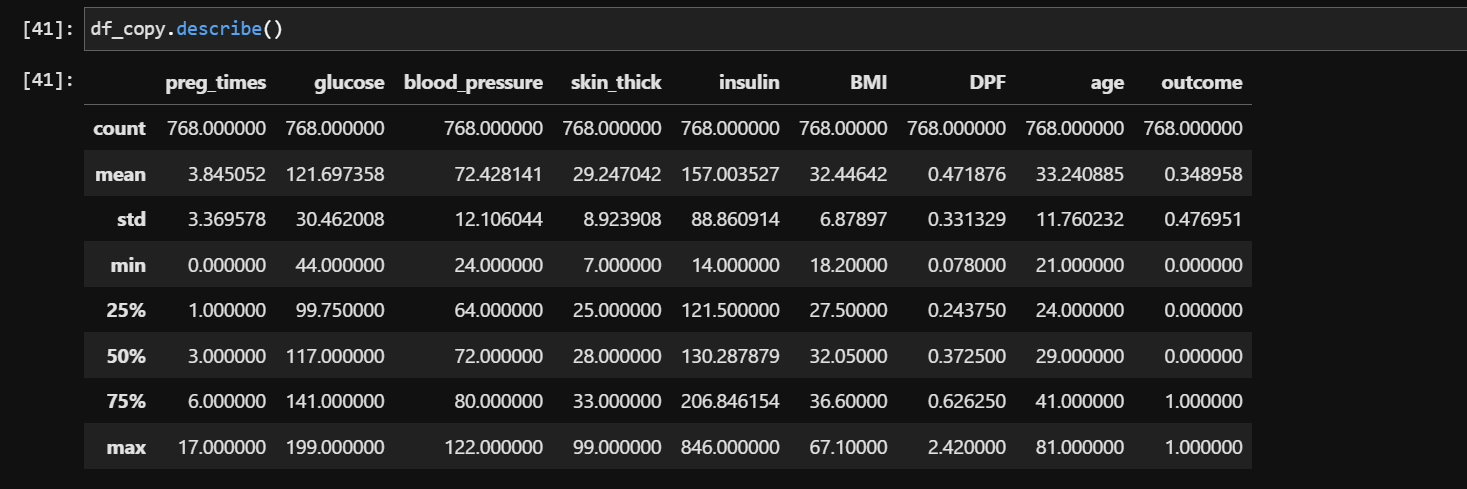
将平均值填入缺失值（以glucose为例，其他类似）



填充完毕后，查看各列缺失值，数据集无缺失值。

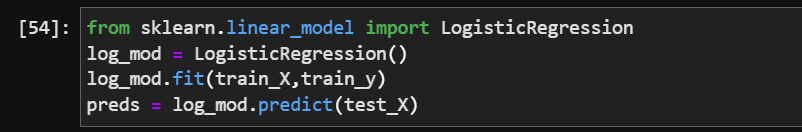


处理后的数据

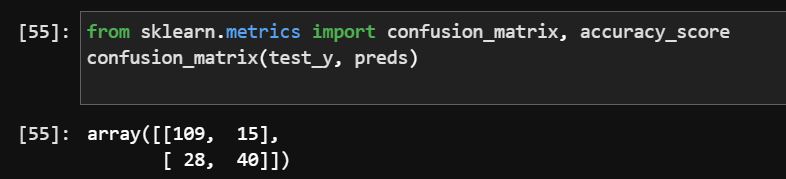


（2）

使用逻辑回归算法训练模型



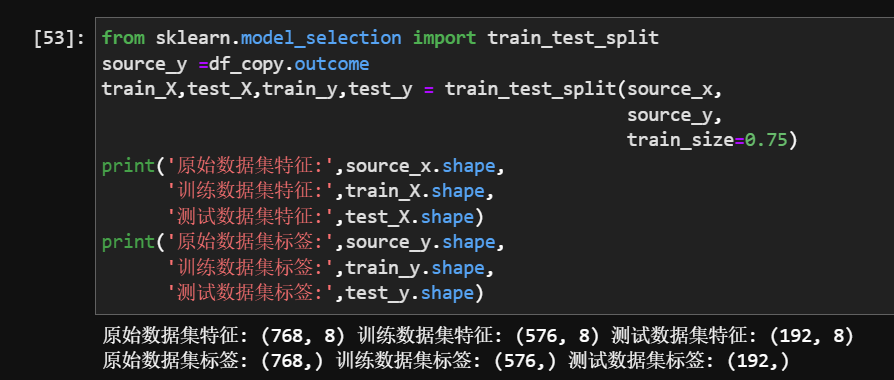
使用混淆矩阵评估模型



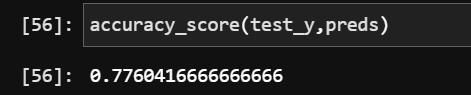
在混淆矩阵中可知，154名皮马印第安女性中，被正确识别为未患糖尿病的正常人数量为109名， 被正确识别为患糖尿病的糖尿病患者数量为40名，被识别为未患糖尿病的糖尿病患者为15名，被识别为糖尿病的正常人为28名。

（3）使用train\_test\_split函数对数据集进行拆分，并在拆分得到的测试集上测试模型准确率。设置train\_size为0.75、0.8、0.85。

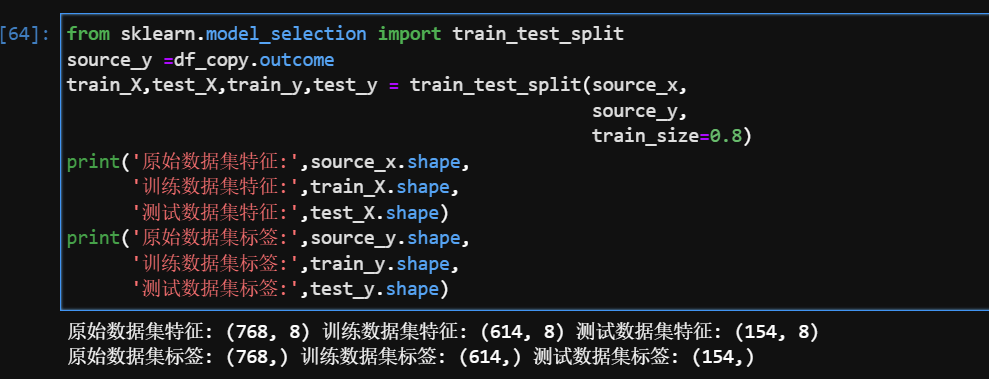
0.75：



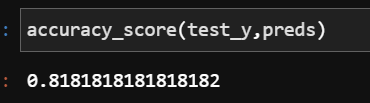
模型准确率：



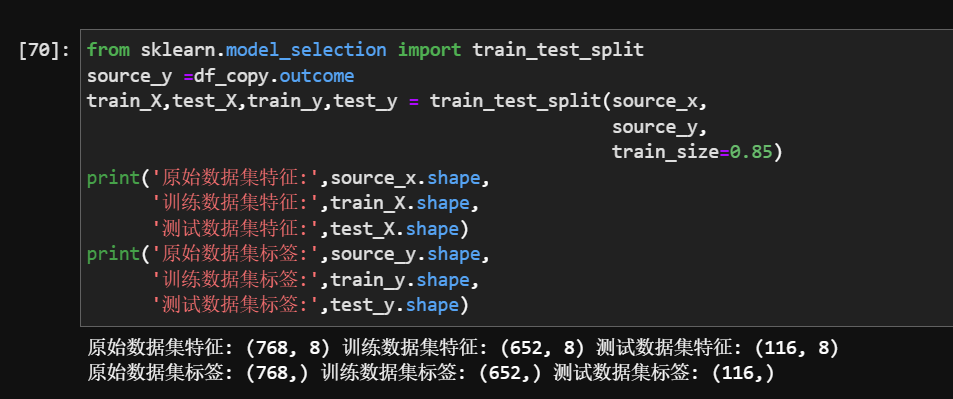
0.8：



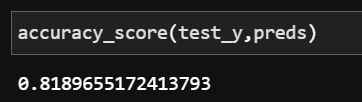
模型准确率：



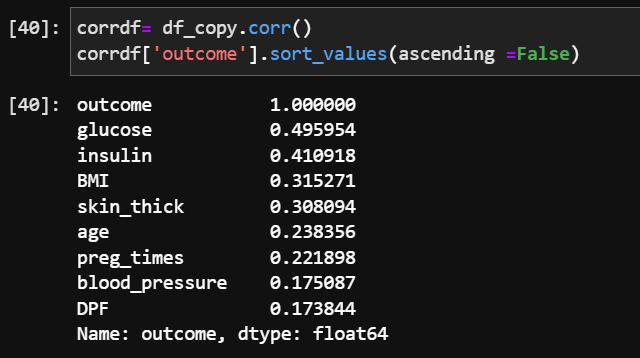
0.85：



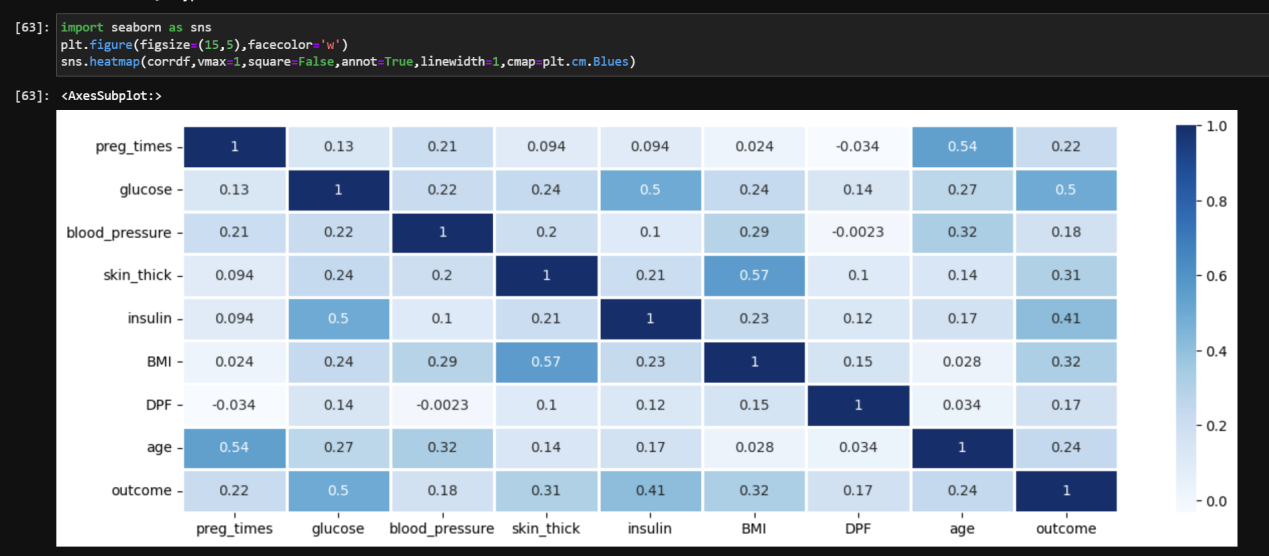
模型准确率：



1. 各个特征值与病情的相关性都较高，且都呈正相关，其中glucose，insulin的相关性最高，blood\_pressure,DPF的相关性最低。



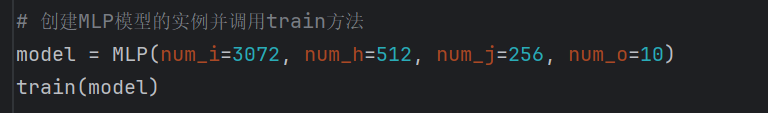
可视化展示：



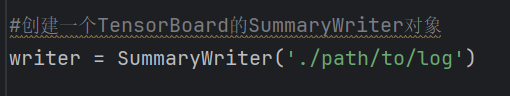
**2.多层感知机：分类问题**

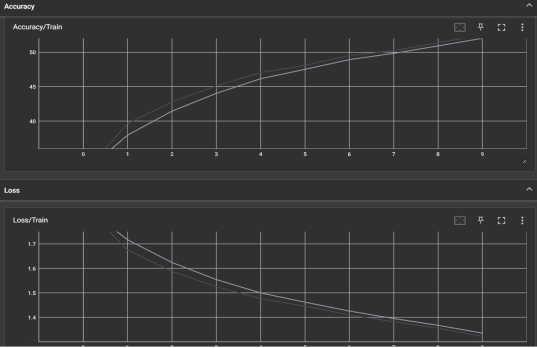
（1）输出网络结构

输入层——>第一隐藏层（3072个神经元）——>第二隐藏层（512个神经元）——>第三隐藏层（256个神经元）——>输出层（10个神经元）

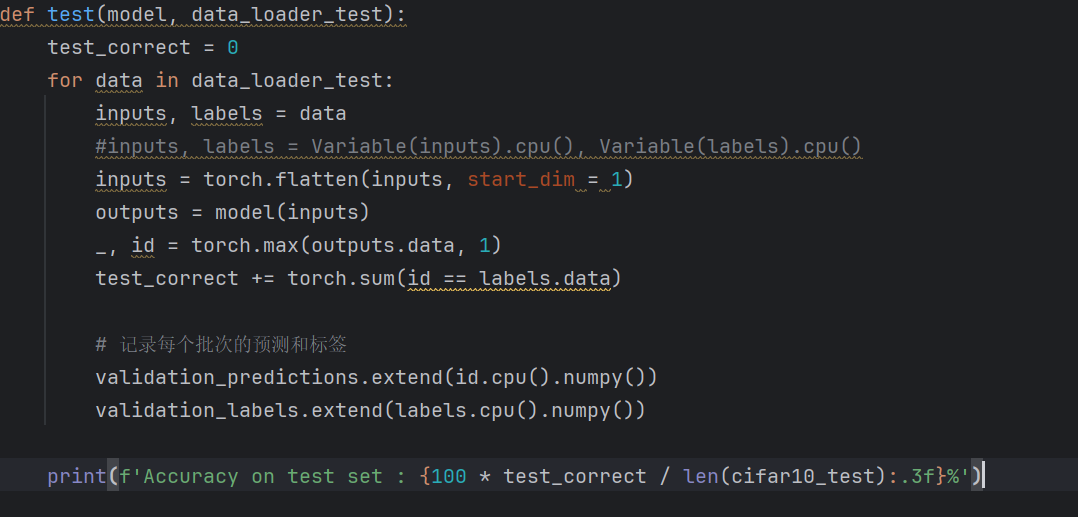


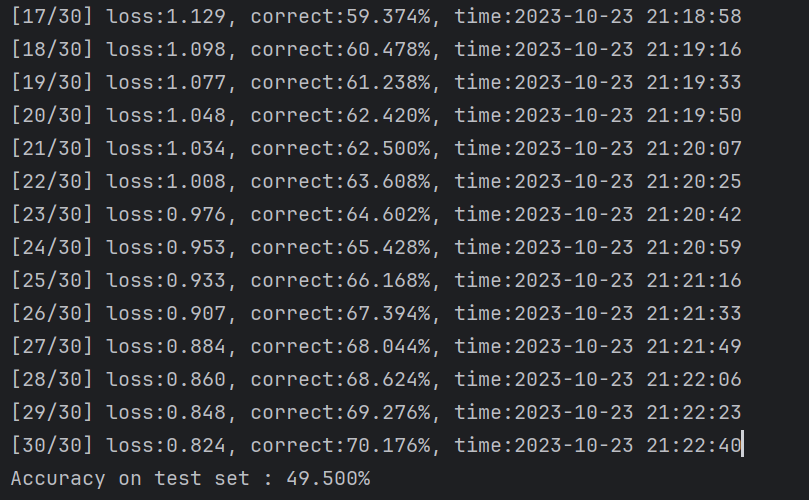
1. 使用tensorboard对训练过程中的loss和accuracy进行可视化展示：



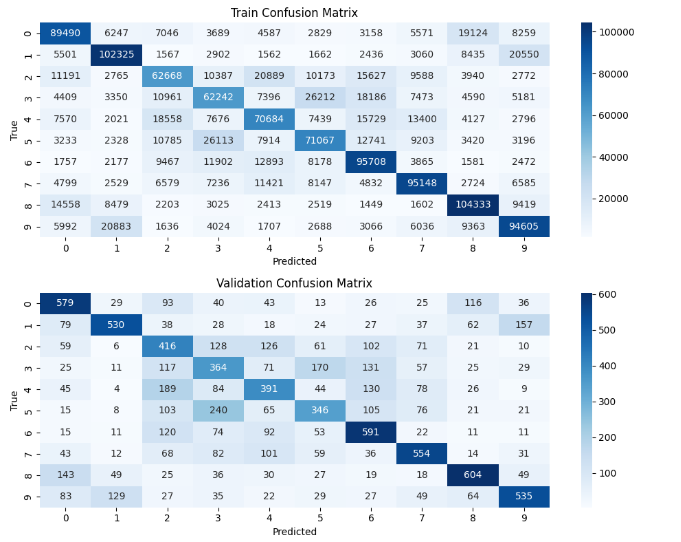


（3）保存训练模型为模型文件，并使用训练好的模型对测试集中的图片进行预测，输出预测结果与预测概率：





（4）训练集和验证集的混淆矩阵：



1. 分析网络参数（例如网络深度、不同的激活函数、神经元数量等）对预测结果的影响。

网络深度：较深的网络可以更好地捕获层次化特征，对于复杂任务，深层网络可能更有优势。然而，过于深的网络可能会导致梯度消失或梯度爆炸问题，因此需要适当的正则化和初始化方法。

在本实验中，网络深度过深时会导致梯度爆炸的问题，因此需要适当降低网络深度。

激活函数：不同的激活函数对于网络的学习能力和训练过程都有重要影响。

ReLU（Rectified Linear Unit）是一种常用的激活函数，它在很多情况下能够提供更好的梯度传播，有助于快速训练。

Sigmoid和Tanh等函数可以用于二元分类问题，但在深层网络中可能出现梯度消失问题。

Leaky ReLU、ELU和Swish等激活函数是改进版本，可以减轻一些传统激活函数的问题。

神经元数量：过多的神经元可能导致过拟合，尤其是在数据集较小的情况下，需要适当的正则化来防止这种情况。在本实验中，神经元不宜过多，应设置适当的值。

（6 ）在损失函数为交叉熵的情况下，对比网络最后一层是否使用softmax的性能差异并分析其产生的原因。

在本实验中，最后一层使用softmax会导致预测准确率降低，性能并不如不使用softmax时的情况。

使用Softmax的情况下：

适用于多分类问题，尤其是具有多个互斥类别的问题，如图像分类。

输出是标准化的概率分布，可以直接用于多类别概率估计和解释。

可以通过交叉熵损失来最小化分类错误。

不使用Softmax的情况下：

适用于二元分类问题，其中只有两个可能的类别。

适用于某些特殊任务，例如目标检测中的边界框回归，其中网络输出可能不是类别概率，而是其他信息，如坐标等。

总之，性能差异主要取决于任务的特性。对于多类别分类问题，Softmax通常是一个更好的选择，因为它提供了标准化的概率分布和相关的损失函数。在其他情况下，根据任务的要求和输出的特性，可以选择不使用Softmax，而是使用适当的激活函数。