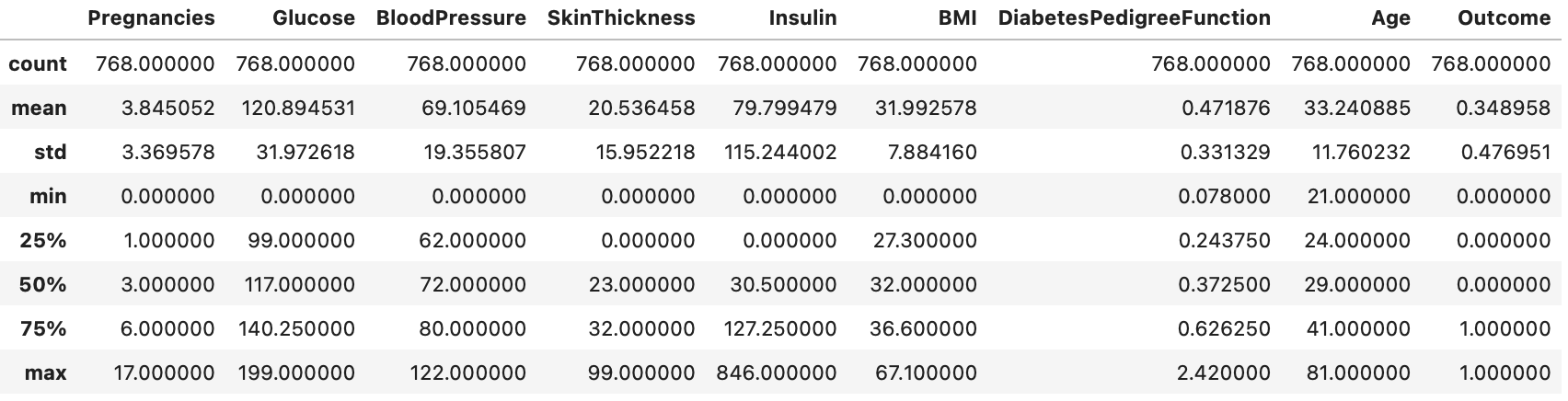
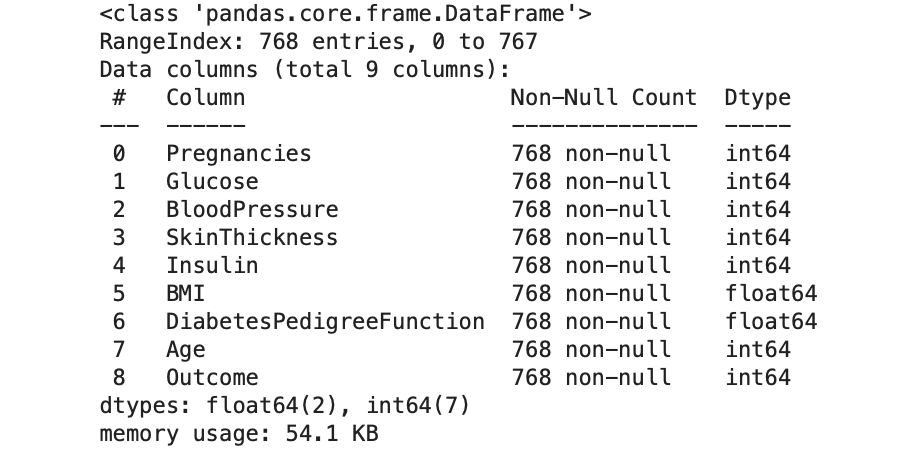
### 糖尿病预测模型（KNN 与SVM）

#### 数据概述与预处理

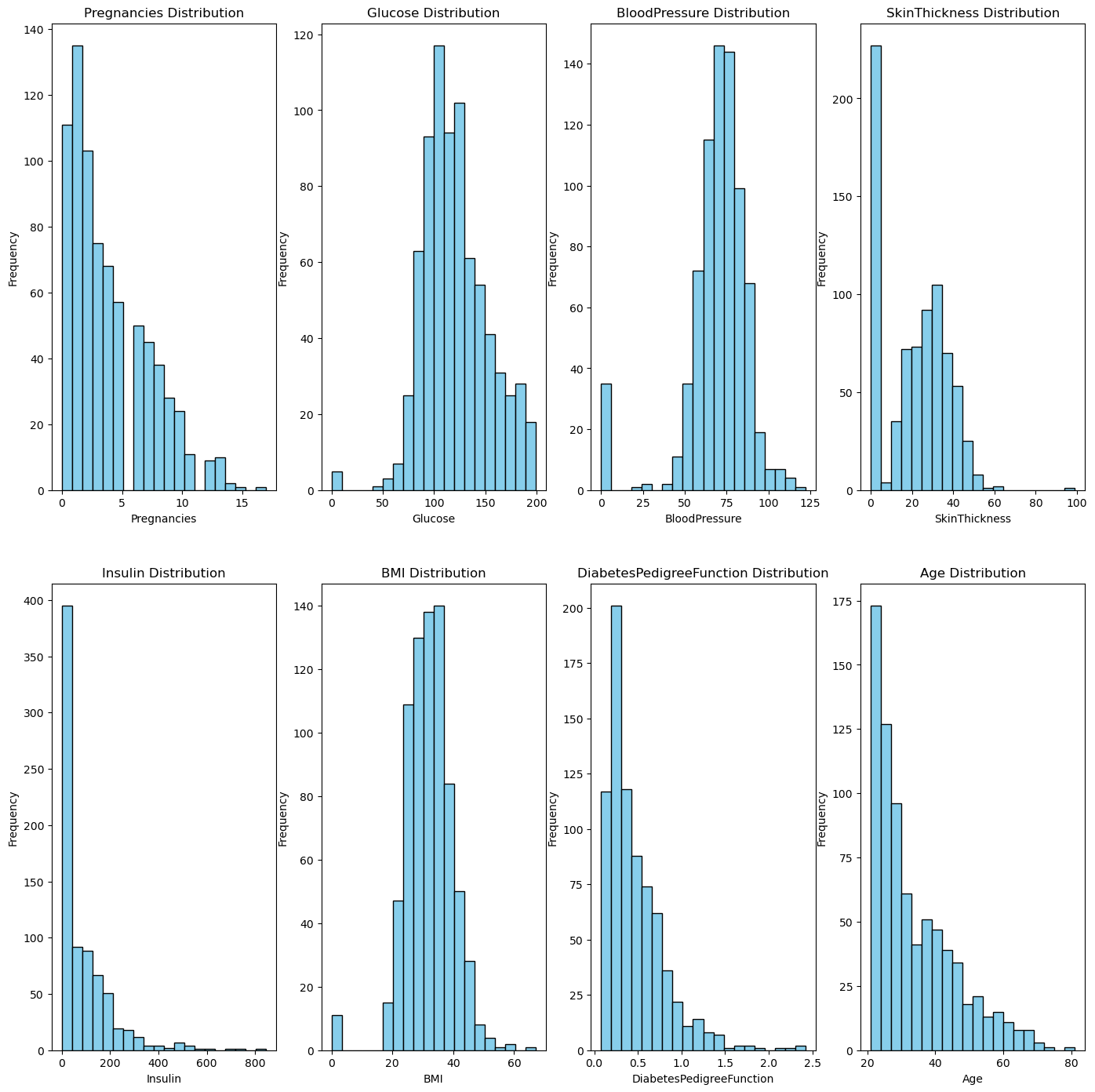
本次分析使用的数据集包含 768 条记录，每条记录代表一个患者的糖尿病相关信息。数据集共有 9 个特征列，包括与患者健康状况和生活习惯相关的重要指标，具体如下：**Pregnancies**（怀孕次数）、**Glucose**（血糖水平）、**BloodPressure**（血压）、**SkinThickness**（皮肤厚度）、**Insulin**（胰岛素水平）、**BMI**（体重指数）、**DiabetesPedigreeFunction**（糖尿病家族史）、**Age**（年龄），以及结果变量 **Outcome**，其中 0 表示未患糖尿病，1 表示患有糖尿病。

通过数据的初步分析，确认了数据集的完整性，所有特征均没有缺失值。为了进一步了解数据分布情况，我们使用了 data.describe() 进行统计描述，得到了各特征的均值、标准差以及最小值和最大值等信息。不同特征数值范围差距较大，可能需要进行标准化，对于方差较大的可能出现异常值。

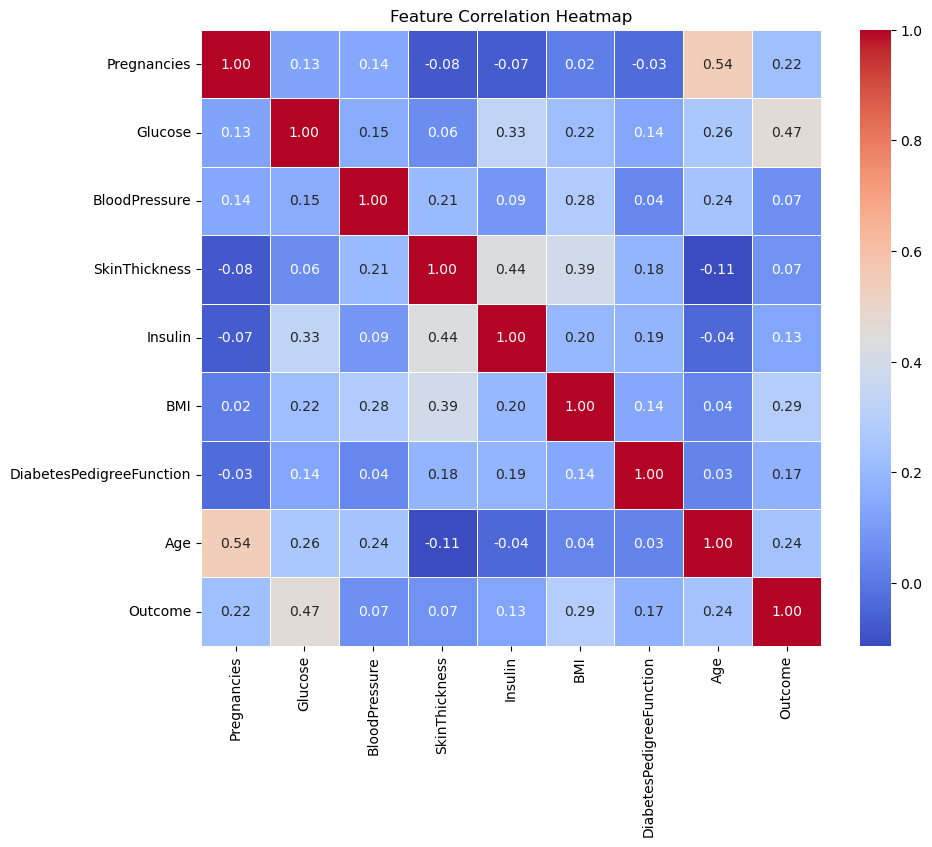


#### 探索性数据分析 (EDA)

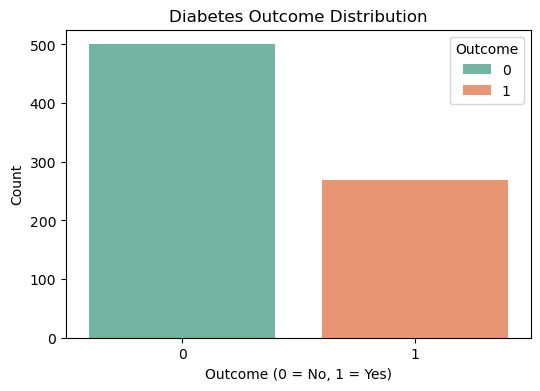
在对数据进行初步探索时，我们通过绘制各特征的分布直方图，发现多个特征呈现右偏分布，特别是 **Pregnancies**、**Insulin**、**DiabetesPedigreeFunction** 和 **Age**。具体来说，**Pregnancies**、**Insulin** 和 **DiabetesPedigreeFunction** 的正常值通常较小，因此它们的分布呈现右偏。而 **Age** 特征的右偏分布可能导致数据不均衡，尤其是在处理糖尿病预测问题时，右偏分布的特征可能使得某些数据点对模型产生不平衡的影响，从而影响模型的训练和预测效果。



接着，通过相关性分析，我们使用热力图展示了各特征之间的相关性，结果显示 **Glucose** 和 **Outcome** 之间存在强相关，符合我们对糖尿病预测的基本认识，同时 **BMI** 和 **Age** 也与糖尿病的发生相关。



在对数据进行分析时，**目标变量**（**Outcome**）的分布也进行了可视化。使用了 **Seaborn** 库的 **countplot** 函数展示了糖尿病的分布情况。结果显示，**Outcome** 中，标签为 0（表示没有糖尿病）和标签为 1（表示有糖尿病）之间的分布不均衡。图中可以清楚地看到，类别 0的样本数量远高于类别 1的样本数量，体现了数据集中的**类别不平衡问题**。这一不平衡可能会对模型的训练产生影响，尤其是在使用 **KNN** 或 **SVM** 等分类算法时，可能导致对少数类（糖尿病）预测能力的下降。



#### 数据清洗与特征选择

由于 **BloodPressure**、**SkinThickness** 和 **Insulin** 与目标变量 **Outcome** 的相关性较低，我们决定在后续建模中剔除这些特征，从而减少模型复杂度并提高计算效率。

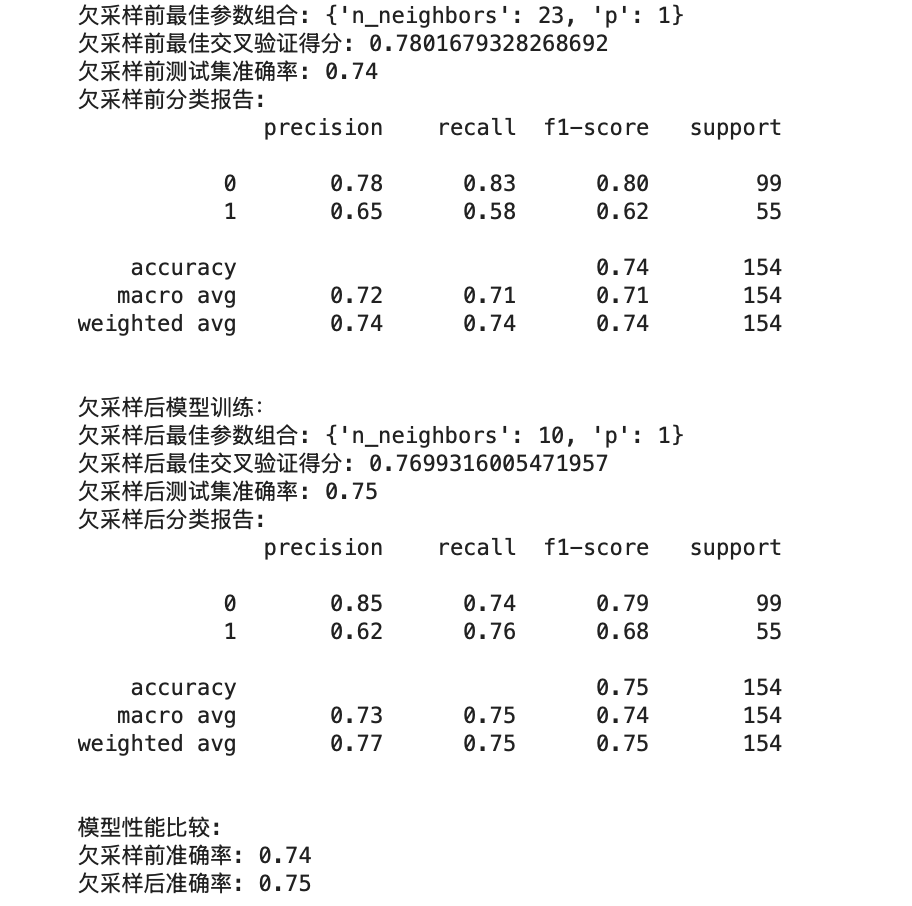
#### 欠采样与模型训练

在模型训练过程中，我们曾尝试使用 **SMOTE（合成少数类过采样技术）** 来解决类别不平衡问题。SMOTE 通过生成新的少数类样本来增加糖尿病患者的数量。然而，使用 SMOTE 后，模型的效果并不好。原因在于，SMOTE 生成的样本可能过于接近现有样本，导致模型容易过拟合。此外，这些合成样本并不总是能够真实反映数据的分布，尤其是在复杂的决策边界情况下。

因此，我们决定转而使用 **欠采样** 方法，即减少非糖尿病患者（类别 0）的样本数量，这样可以更好地平衡数据集，并提高模型的性能。

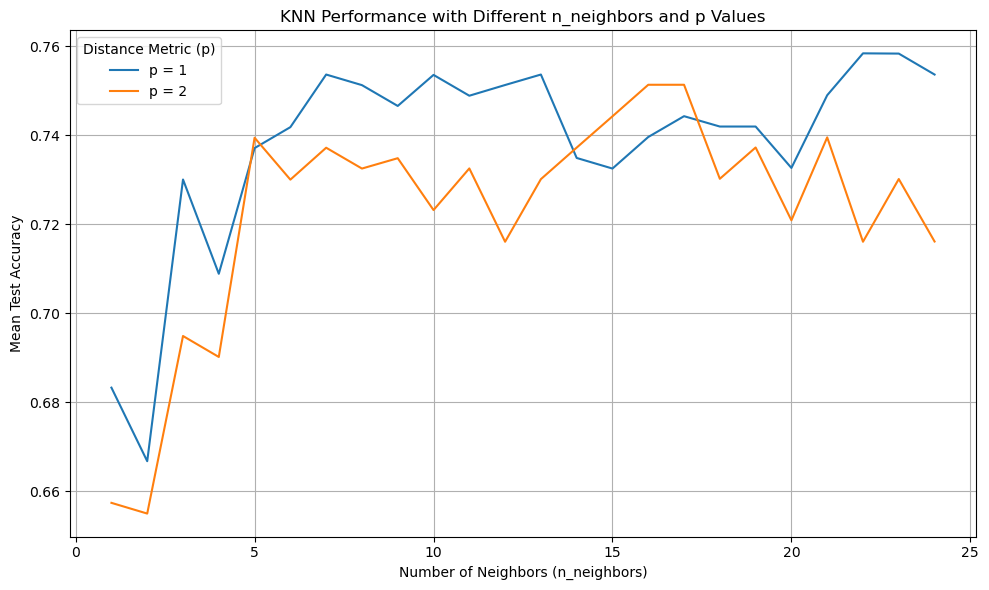
我们选择了 **KNN（K-最近邻）** 和 **SVM（支持向量机）** 两个模型进行训练，并对欠采样前后的模型性能进行了比较。

##### KNN 模型



在欠采样前，KNN 模型的最佳参数为 n\_neighbors=23，p=1（曼哈顿距离），测试集准确率为 74%。而在欠采样后，KNN 模型的最佳参数为 n\_neighbors=10，p=1，测试集准确率提升至75%，而且糖尿病类别的召回率显著提高，表明欠采样有助于提高对少数类（糖尿病）的识别能力。

下面展示了不同参数下的性能，可以看到随着 **n\_neighbors**（邻居数）增加，准确率先是上升的，随后在一个小范围内波动。这是因为，随着 **n\_neighbors** 的增加，KNN 模型开始考虑更多的邻居，模型的决策变得更加平滑。较小的 **n\_neighbors** 值可能导致模型对噪声较为敏感，而过高的 **n\_neighbors** 值则可能使得模型过于简单，无法捕捉到数据中的复杂模式。因此，当 **n\_neighbors** 值过高时，模型的性能可能会受到影响，导致准确率趋于平稳或略有下降。



##### SVM 模型截屏2024-12-07 14.08.07

欠采样前后，SVM 模型的最佳参数选择一致，均为 {'C': 100, 'gamma': 'scale', 'kernel': 'rbf'}。欠采样前，模型的准确率为 0.76，表现较好，特别是在区分非糖尿病（类别 0）时，召回率为 0.85，精确率为 0.79。然而，在糖尿病（类别 1）的识别上，召回率较低，仅为 0.60，精确率为 0.69，表现较差。欠采样后，虽然糖尿病的召回率提高至 0.82，但整体准确率下降至 0.73，同时非糖尿病的召回率下降至 0.68，精确率提高至 0.87，导致精确率和召回率之间的不平衡。综上所述，欠采样改善了糖尿病类别的召回率，但也导致了非糖尿病类别的召回率下降，整体准确率有所下降。

#### 模型性能比较

通过对比 KNN 和 SVM 模型，我们发现 **SVM** 模型在未进行欠采样的情况下表现较好，尤其在区分非糖尿病（类别 0）时，具有较高的召回率和精确率。相对而言，KNN 模型虽然准确率较低，但在改善糖尿病类别（类别 1）的识别上有一定优势。对于欠采样的效果，**KNN** 模型表现出更显著的改进，尤其是在提高糖尿病类别的召回率方面。而 **SVM** 模型在欠采样后虽然糖尿病的召回率有所提升，但整体准确率却有所下降，表明欠采样在某些情况下可能导致对非糖尿病类别的识别能力下降。与 KNN 模型相比，SVM 的效果更好。这是因为 SVM 能更好地拟合高维数据，并且其 **rbf** 核函数能够更精确地捕捉数据的决策边界。SVM 在处理高维特征空间时，能够通过非线性映射找到更合适的决策边界，从而提升分类性能，尤其在数据集具有复杂边界时，SVM 的表现优于 KNN。

#### 结论与建议

通过本次分析，发现欠采样方法对 KNN 和 SVM 模型在糖尿病预测任务中的表现有一定影响。对于 KNN 模型，欠采样有助于提高糖尿病类别的召回率，尽管整体准确率略有下降。而对于 SVM 模型，欠采样可能会导致整体性能下降，尤其是在识别非糖尿病患者时。因此，在处理数据不平衡问题时，需要综合考虑精确率和召回率之间的平衡，以便优化模型的整体性能。