目录

第一章 问题背景与提出

1.1. 问题背景

糖尿病，这一在全球范围内普遍流行的慢性非传染性疾病，不仅患者人数众多，而且对广大民众的健康构成了严峻挑战，同时给公共卫生系统施加了巨大压力。鉴于此，精准评估糖尿病的发病风险对于实施早期干预措施及策划高效的预防策略显得尤为重要，具有无可估量的价值和深远的战略意义。

1.2. 问题提出

我们期望借助现有的数据集，构建出一个模型，该模型能够精确识别出驱动糖尿病发病的核心因素，并有效预测糖尿病的发病风险，从而为医学决策提供有力支持，辅助专业人士做出更为精准的判断。

第二章 项目框架

本项目首先运用数据可视化技术对相关数据进行初步的探索与分析，以直观地了解数据的分布与特征。接着，采用斯皮尔曼相关性检验深入挖掘影响糖尿病发病的各类因素，探究各因素与糖尿病之间的关联程度。随后，构建逻辑回归模型、随机森林模型与 XGBoost 预测模型，通过对比它们在预测患病风险方面的性能表现，选取更为合适的模型，并对其进行进一步的优化调整，深度剖析模型中的重要特征，精准识别出影响糖尿病发病的关键因素，从而为糖尿病的早期干预和预防工作提供强有力的科学依据与决策支持，助力提升糖尿病防治工作的成效与水平。

项目的 Guihub 链接为：https://github.com/Mufan0502/Machine-learning

第三章 数据探索

3.1. 数据准备

数据来源：

https://www.kaggle.com/datasets/rabieelkharoua/diabetes-health-dataset-analysis/code

数据集中一共有 1879 行记录，包含有 46 个特征（如患者 ID、年龄、性别、是否患糖尿病等等）。

3.2. 数据概况

数据集中 46 个特征，描述性统计信息如下：

首先通过 info()函数查看特征类型以及有多少非空值。从下图可以看出，所有特征都有 1879 个非空值，说明该数据集不存在数据缺失。此外，除了最后一个特征'DoctorInCharge'（主治医生）为字符串型数据，其余特征均是数值型。

PatientID	Age	Gender	Ethnicity	EducationLevel	...	WaterQuality	MedicationAdherence	HealthLiteracy	Diagnosis	DoctorInCharge
1879 non-null	1879 non-null	1879 non-null	1879 non-null	1879 non-null	...	1879 non-null	1879 non-null	1879 non-null	1879 non-null	1879 non-null
int64	int64	int64	int64	int64	...	int64	float64	float64	int64	object

接下来通过 describe()函数查看统计信息，如下图：

	PatientID	Age	Gender	Ethnicity	EducationLevel	...	WaterQuality	MedicationAdherence	HealthLiteracy	Diagnosis	DoctorInCharge
count	1879	1879	1879	1879	1879	...	1879	1879	1879	1879	1879
unique	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	1
top	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	Confidential
freq	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	1879
mean	6939	55.04	0.487493348	0.755721	1.699308143	...	0.200638638	4.957539393	5.01173646	0.400212879	NaN
std	542.5648963	20.52	0.49997662	1.047558	0.885665367	...	0.400584792	2.910933516	2.920908394	0.490071779	NaN
min	6000	20	0	0	0	...	0	0.00538376	0.000362249	0	NaN
25%	6469.5	38	0	0	1	...	0	2.420023686	2.410112691	0	NaN
50%	6939	55	0	0	2	...	0	4.843885844	5.035208098	0	NaN
75%	7408.5	73	1	1	2	...	0	7.513933228	7.58686466	1	NaN
max	7878	90	1	3	3	...	1	9.997164754	9.993029038	1	NaN

第一个特征 'PatientID' 作为患者的唯一标识符，对于预测糖尿病无直接贡献；特征 'DoctorInCharge'（主治医生）只有一个唯一值，而且是 Confidential（保密），所以这一列特征也没办法提供更多的有用信息。综上，可以将这两个字段从特征集中剔除，其他特征无异常。

3.3. 特征分布可视化

去掉目标变量'Diagnosis'，剩余的特征可以分为八大类：

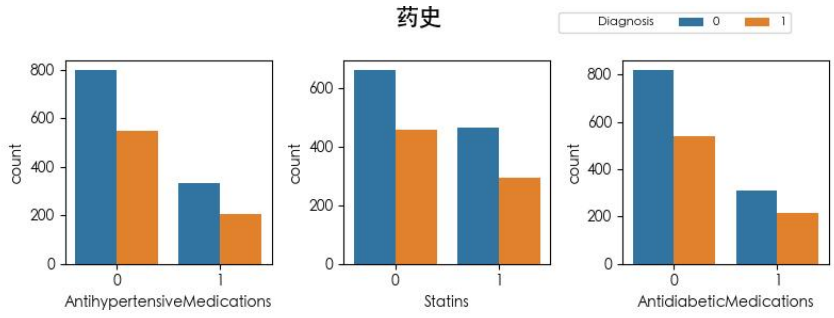
- 药史，包含：'AntihypertensiveMedications'，'Statins'，'AntidiabeticMedications'；
- 个人统计信息，包含：'Age'，'Gender'等共 6 个特征；
- 生活方式，包含：'Smoking'，'AlcoholConsumption'等共 6 个特征；
- 病史，包含 5 个特征；
- 生活环境，包含：'HeavyMetalsExposure'，'OccupationalExposureChemicals'，'WaterQual9ity'；
- 健康情况，包含：'MedicalCheckupsFrequency'，'MedicationAdherence'，'HealthLiteracy'。
- 医学测量结果，包含：'SystolicBP'，'FastingBloodSugar'等 10 个特征；
- 症状，包含：'FrequentUrination'，'ExcessiveThirst'，'UnexplainedWeightLoss'，'FatigueLevels'，'BlurredVision'，'SlowHealingSores'，'TinglingHandsFeet'；

接下来，我们通过可视化来初步探究特征分布与是否患糖尿病的关系。

分类别去看，各个类别的特征分布可视化结果如下：

3.3.1. 药史

从左到右分别为降压药使用情况、他汀类药物使用情况、降糖药物使用情况在不同诊断结果下的分布情况。0 表示不患糖尿病，1 则是患病。

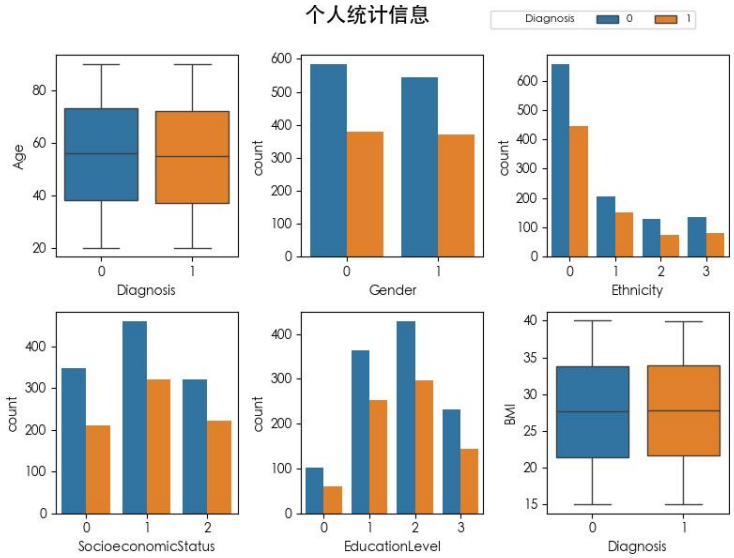


可以看出，诊断结果为 0 和 1 的患者在药物使用情况上的分布趋势相似。这可能

暗示这些药物使用情况与诊断结果之间的关系不太显著

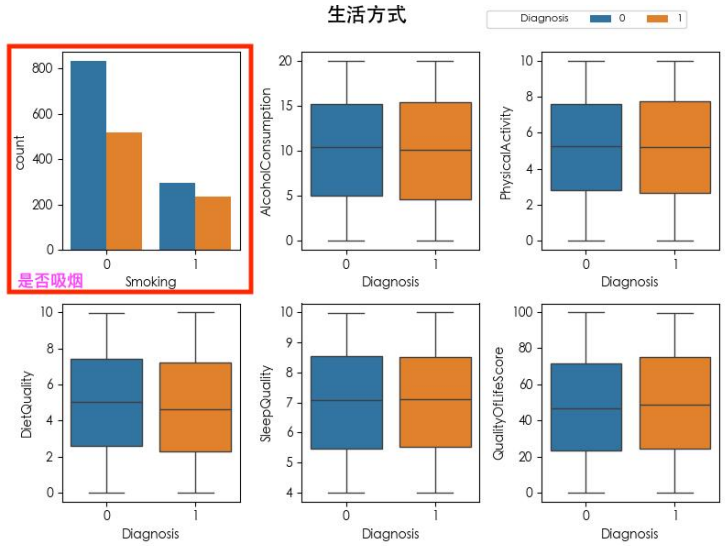
3.3.2. 个人统计信息

下面展示了个人统计信息特征与诊断结果之间的关系。通过观察这些图表，初步判断个人统计信息，包含年龄、性别等 6 个特征，对诊断结果的影响不大。



3.3.3. 生活方式

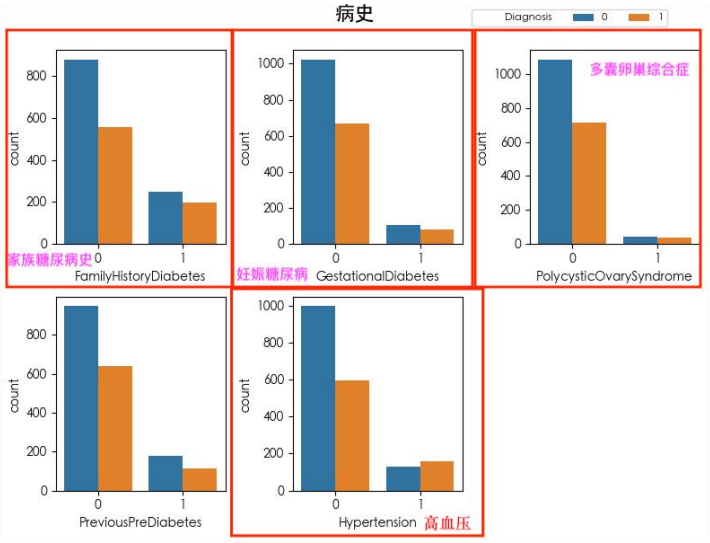
吸烟的柱状图显示，吸烟（横坐标为 1）的患者中，诊断结果为 1（橙色）的患者占比明显高于不吸烟（横坐标为 0）的患者中患糖尿病的占比。这表明吸烟可能与确诊患病相关。其余特征在不同诊断结果中的分布相似，可能对诊断结果的影响不大。



3.3.4. 病史

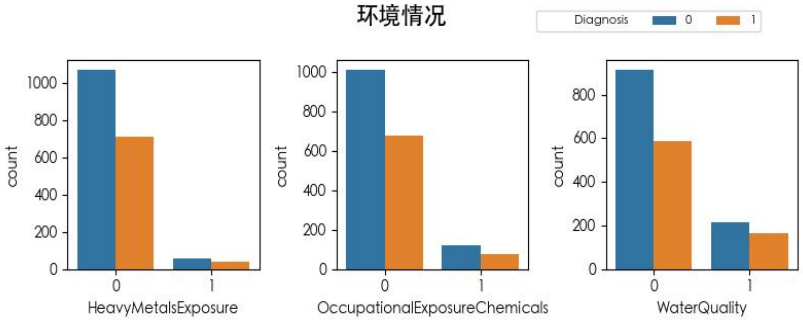
下图展示了多种病史特征与诊断结果之间的关系。通过观察这些图表，我们可以进行以下结论：具有家族糖尿病史、妊娠糖尿病史、多囊卵巢综合症以及高血压病史的患者，其罹患糖尿病的风险显著增加（诊断结果倾向为 1），表明这些病史特征可能与提升患病风险存在关联性。尤其值得注意的是，高血压在这些因素

中表现最为突出：在患有高血压的人群中，糖尿病患者的比例超过了非糖尿病患者；而在无高血压病史的人群中，情况则恰好相反。



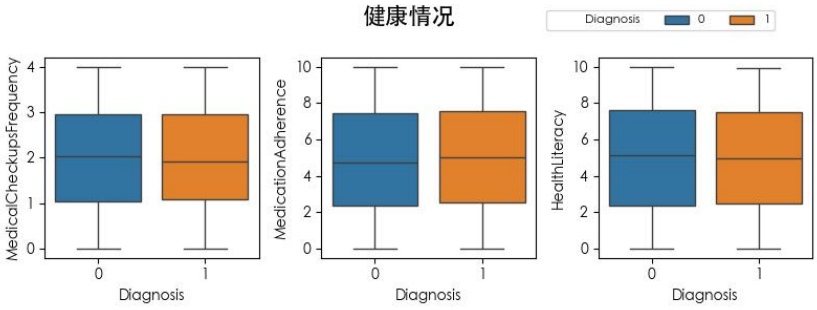
3.3.5. 生活环境

可视化结果如下，可看出，这些特征的分布与诊断结果无显著关系。



3.3.6. 健康情况

同样，健康情况的特征可能对诊断结果也无贡献。



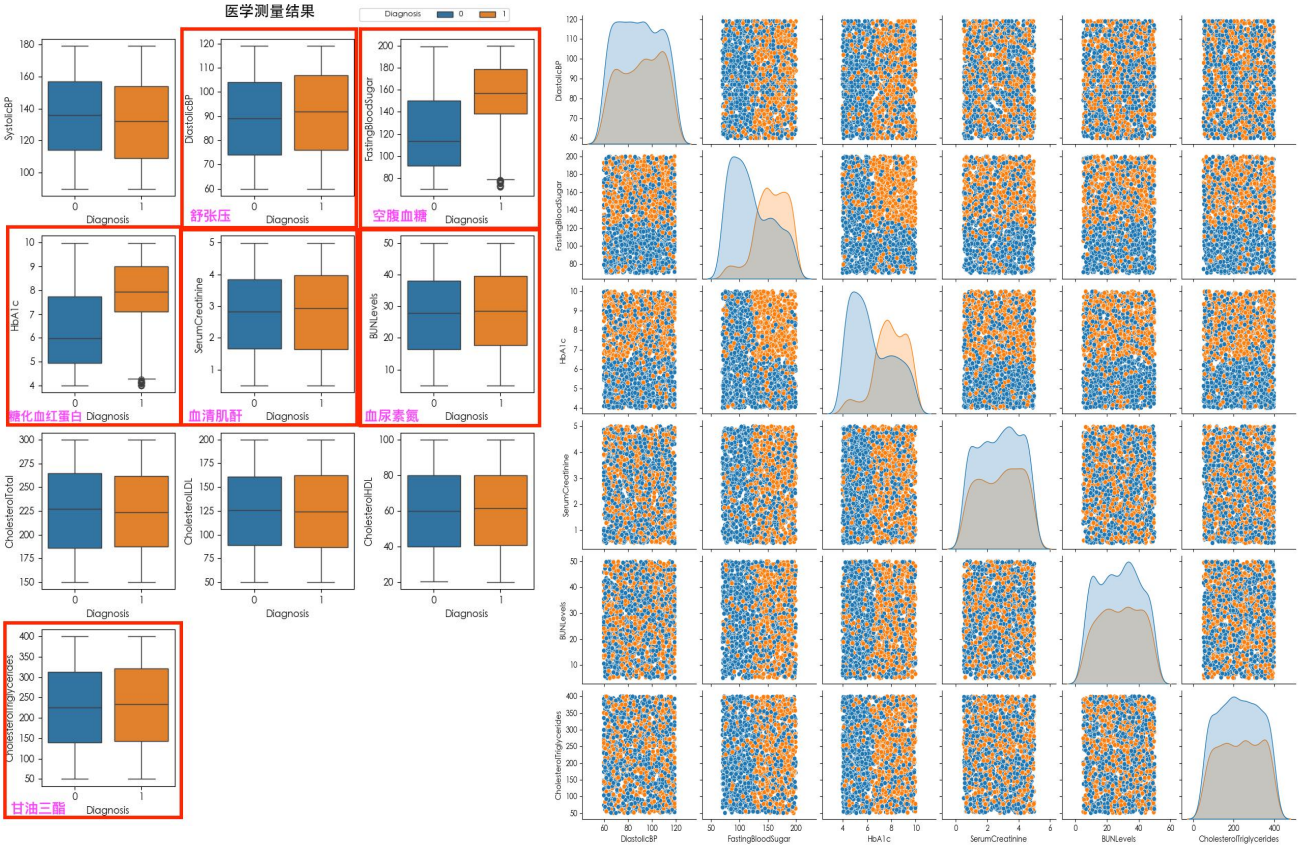
3.3.7. 医学测量结果

从箱线图（左图）来看：

- 空腹血糖、糖化血红蛋白：这两个特征在不同诊断结果中的分布存在显著差异。总体来看，这些特征较高的患者更可能有较高的诊断结果(1)，即这些特征可能与较高的患病风险相关。
- 舒张压、血清肌酐、血尿素氮、甘油三酯：诊断结果为 1 的患者在这几个特征上的分布总位数略高于诊断结果为 0 的患者。

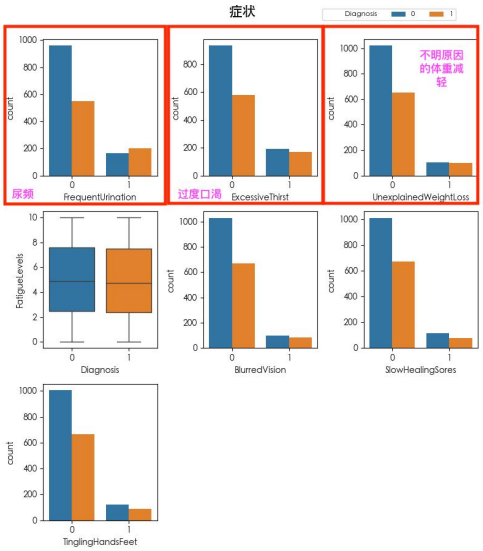
- 其余特征在不同诊断结果中的分布相似。

对箱线图初步分析得到的重要特征绘制成对关系图，得到下面右图。可以看出，空腹血糖与糖化血红蛋白这两个变量的分布图显示出不同诊断结果之间的明显差异，诊断结果为 1 对应这两个指标均偏高，呈现明显的右偏分布。其余特征在不同诊断结果下分布相对均匀。



3.3.8. 症状

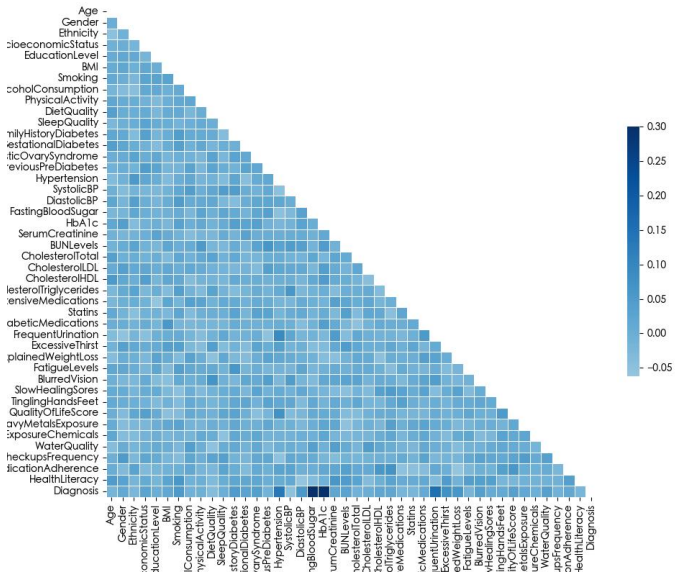
尿频、过度口渴、体重无故减轻三种症状，在两种诊断结果之间的分布差异明显。这暗示这些症状与较高的发病概率有密切联系。



3.4. 斯皮尔曼相关性分析

计算特征与诊断结果'Diagnosis'的斯皮尔曼相关性矩阵，结果如下：

特征	相关性	特征	相关性
Age	-0.016863667	CholesterolTotal	-0.010600631
Gender	0.011746025	CholesterolLDL	-0.000659955
Ethnicity	-0.023132151	CholesterolHDL	0.017333947
SocioeconomicStatus	0.025572128	CholesterolTriglycerides	0.018175383
EducationLevel	-0.002306051	AntihypertensiveMedications	-0.02143288
BMI	0.014585877	Statins	-0.021609559
Smoking	0.053829499	AntidiabeticMedications	0.01136382
AlcoholConsumption	-0.00967076	FrequentUrination	0.151504568
PhysicalActivity	-0.006412725	ExcessiveThirst	0.073524779
DietQuality	-0.041431809	UnexplainedWeightLoss	0.061038454
SleepQuality	-0.002937501	FatigueLevels	-0.017722593
FamilyHistoryDiabetes	0.047681093	BlurredVision	0.038339293
GestationalDiabetes	0.029609786	SlowHealingSores	0.006293513
PolycysticOvarySyndrome	0.038802975	TinglingHandsFeet	0.011592653
PreviousPreDiabetes	-0.011024784	QualityOfLifeScore	0.035148029
Hypertension	0.131883465	HeavyMetalsExposure	0.013577537
SystolicBP	-0.051795623	OccupationalExposureChemicals	-0.005859259
DiastolicBP	0.055145171	WaterQuality	0.032873332
FastingBloodSugar	0.451990385	MedicalCheckupsFrequency	-0.009597619
HbA1c	0.431109478	MedicationAdherence	0.015927185
SerumCreatinine	0.022516389	HealthLiteracy	-0.021669975
BUNLevels	0.028276768		



可以看出：空腹血糖（FastingBloodSugar）与糖化血红蛋白（HbA1c）呈现出最为显著的相关性。其次，尿频（FrequentUrination）、高血压（Hypertension）、过度口渴（ExcessiveThirst）以及不明原因的体重减轻（UnexplainedWeightLoss）等因素与糖尿病的关联程度较为明显。而在这些因素之后，舒张压（DiastolicBP）和吸烟状况（Smoking）与糖尿病也存在一定的相关性，但相对较弱。与直观观察特征分布与糖尿病的关系所得结论基本一致。。

第四章 数据预处理

4.1. 特征选择

弱相关的特征（与糖尿病相关性绝对值小于 0.01）：

弱相关的特征	相关性
EducationLevel	-0.002306
AlcoholConsumption	-0.009671
PhysicalActivity	-0.006413
SleepQuality	-0.002938
CholesterolLDL	-0.00066
SlowHealingSores	0.006294
OccupationalExposureChemicals	-0.005859
MedicalCheckupsFrequency	-0.009598

弱相关特征会对模型造成干扰，因此在后续模型训练中剔除掉这部分特征

4.2. 数据标准化

特征可分为数值型和类别型，数值型特征例如年龄、空腹血糖等等，类别型特征有性别、种族等等。对于数值型特征，我们采取标准化处理手段，以统一其尺度。

具体公式如下：

$$X' = \frac{X - mean}{\sigma}$$

4.3. 数据分割

使用 train_test_split 函数将数据集分割为训练集和测试集，其中测试集占总数据集的 30%，并设置随机数种子为 42 以确保结果的可重复性。

第五章 模型训练与优化

5.1. 模型建立

在本次机器学习任务中，我们致力于通过比较和评估不同的模型，以识别出在处理该数据集时表现最优的算法。本次评估聚焦于三种主流模型：逻辑回归(Logistic Regression）、随机森林（Random Forest）以及 XGBoost。这些模型的选择基于其各自的独特优势与适用场景：

- 逻辑回归：我们的最终目标是预测个体是否患有糖尿病，本质上是一个二分类问题，即预测 0 或 1 的变量，而逻辑回归通常用于处理二分类问题，且其简单易懂、易于解释的特性，使得特征的重要性分析成为可能。
- 随机森林：随机森林作为一种集成学习方法，通过构建多个决策树并综合其预测结果，实现了模型准确性和鲁棒性的显著提升。在处理包含高维数据、噪声数据及异常值的复杂场景时，随机森林展现出了非凡的能力。特别地，对于特征数量众多（如我们考虑的数据集包含 40 多个特征）且特征间可能存在非线性关系的情况，随机森林尤为适用。
- XGBoost：XGBoost 是一种基于梯度提升的集成学习方法，以其高效性、灵活性和可扩展性得到广泛应用。在处理大规模数据集和构建复杂模型的任务中，XGBoost 展现出了优越的性能。它能够有效地处理数据中的缺失值，非常适合处理特征数量较多且存在复杂非线性关系的数据集。考虑到我们的数据集特征丰富，且存在正负样本比例约为 2:3 的不平衡问题，XGBoost 的这些特性使其成为我们分析中的一个重要选项。

5.2. 模型调参

为了优化每个模型的性能，我们使用了网格搜索（Grid Search）和交叉验证（Cross-Validation）在指定的参数范围内自动寻找最佳模型参数组合，以提升机器学习模型在测试集上的准确性，并输出最佳参数及相应的模型准确性评估结果。具体的调参过程如下：

网格搜索会遍历所有候选的参数组合，而 5 折交叉验证则通过将训练数据集划分为五个子集，轮流使用其中四个子集进行模型训练，剩余一个子集用于模型验证，以此方式来评估每个参数组合的性能。这一过程确保了参数选择的客观性和稳定性。最终，该方法会输出在验证过程中表现最佳的参数组合，以及该参数

组合下模型在测试集上的准确性评估结果，为模型调优提供了有力的数据支持。

- 逻辑回归:

在逻辑回归模型中，我们调整了正则化系数（predictor__C，取值范围为{0.1, 1, 10, 20}）和最大迭代次数（predictor__max_iter，取值范围为{100, 200, 300}）。正则化系数 C 则用于调控模型的复杂度，旨在避免模型出现过拟合现象，而最大迭代次数决定了模型训练过程的收敛速度和稳定性，可以找到在保证模型性能的同时，训练过程的计算成本也相对合理。

- 随机森林

对于随机森林模型，我们主要探索了决策树的数量（predictor__n_estimators 的取值范围从 1 至 19）以及决策树的最大深度（predictor__max_depth 的取值范围从 1 至 19）。决策树的数量对模型的稳定性和泛化能力具有显著影响，而决策树的最大深度则限制了模型的复杂度，有助于防止模型因过度拟合而失去泛化能力。通过对这些参数的细致调整，我们可以寻找到在给定数据集上表现最优的随机森林模型配置。

- XGBoost

在 XGBoost 模型中，我们重点调整了决策树的最大深度（predictor__max_depth 的取值范围从 1 至 19）以及最小子节点权重（predictor__min_child_weight 的取值范围从 1 至 19）。决策树的最大深度是控制模型复杂度、防止过拟合的关键因素之一，而最小子节点权重则影响了分裂决策的制定，同样对模型的过拟合控制具有重要影响。此外，我们还根据训练数据中正负样本的比例，在模型初始化时设置了 scale_pos_weight 参数，它为训练集中负类（y_train == 0）与正类（y_train == 1）的比值，以应对数据集不平衡的问题，从而进一步提升模型的性能。

5.3. 模型选择与分析

5.3.1. 模型选择

在模型评估过程中，我们使用了各模型经过网格调参后在测试集上的准确率（accuracy）作为评估指标。以下是每个模型对应的准确率:

模型	准确率
逻辑回归	0.839
随机森林	0.906
XGBoost	0.943

从以上结果可以看出，XGBoost 模型在测试集上的表现最佳，达到了 94.3%的准确率。这表明 XGBoost 在处理当前数据集时，能够更好地捕捉数据中的复杂模式和特征关系。因此我们选择 XGBoost 作为最终的预测模型。

5.3.2. 原因分析

在本次实验中，XGBoost 模型在测试集上的表现优于逻辑回归和随机森林，

达到了 94.3%的准确率。以下是一些解释 XGBoost 优于其他两个模型的原因：

- 数据不平衡处理

该数据集中正类与负类的比例约为 2:3，存在数据不平衡的现象。如下：

```
print(f"样本中被诊断为糖尿病的有{sum(df['Diagnosis']==1)}人，没有糖尿病的有{sum(df['Diagnosis']==0)}人")
```

```
样本中被诊断为糖尿病的有752人，没有糖尿病的有1127人
```

数据不平衡会导致模型在训练时偏向于多数类，从而对少数类的预测效果不佳。这种现象在医学诊断任务中尤为常见，因为真正患病的个体往往远少于未患病的个体。

XGBoost 通过‘scale_pos_weight’参数调节正负样本的权重，使得模型能够更好地平衡正负样本，从而提高对少数类的预测准确率。相比之下，虽然随机森林通过构建多个决策树并综合其预测结果来提高模型的鲁棒性和准确性，但在面对不平衡数据集时，单个决策树可能仍然会偏向于多数类。逻辑回归作为广义线性模型，在处理不平衡数据集时，模型的决策边界可能会偏向于多数类。

- 梯度提升算法的优势

XGBoost 基于梯度提升算法，通过逐步构建和优化树模型来提高预测性能。每一棵新树都是在前一棵树的基础上进行优化，重点关注前一棵树预测错误的样本。这种逐步优化的过程使得 XGBoost 能够更好地捕捉数据中的复杂模式和非线性关系，从而提高模型的准确性。而随机森林虽然通过集成多个决策树来提高模型的鲁棒性，但每棵树的构建是独立的，无法像 XGBoost 那样逐步优化。在糖尿病预测任务中，特征之间可能存在复杂的非线性关系，逻辑回归无法有效捕捉这些关系，从而影响预测性能。

- 正则化和剪枝

XGBoost 具有强大的正则化功能，通过对树模型进行正则化处理（包括 L1 和 L2 正则化），可以有效防止过拟合。此外，XGBoost 还采用了剪枝技术，通过剪枝来去除不必要的分支，进一步提高模型的泛化能力。相比之下，逻辑回归虽然也可以进行正则化处理，但其模型复杂度较低，无法捕捉复杂的非线性关系。而随机森林虽然通过随机采样和特征选择来防止过拟合，但在处理高维数据和复杂关系时，不如 XGBoost 高效。

综上所述，XGBoost 在处理不平衡数据集、捕捉复杂非线性关系、正则化和剪枝、高效计算等方面具有显著优势。这些优势使得 XGBoost 在本次实验中表现优于逻辑回归和随机森林，成为最佳的预测模型。

5.4. 结果分析

对出来相对更优的 XGBoost 模型在该数据集上的表现进行评估与分析。

5.4.1. 模型评估

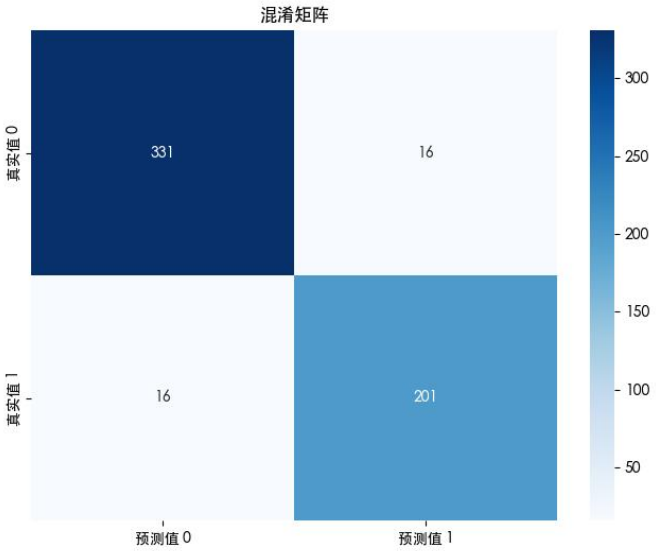
- 分类报告

模型的分类报告包含精确率(Precision)、召回率(Recall)、F1 分数(F1-score)、支持(support)、准确率(Accuracy)、宏平均(Macro avg)、加权平均(Weight avg)。

精确率和召回率分别衡量了模型在预测正类样本时的准确性和覆盖率，F1 分数则综合了这两者的表现。支持指标显示了每个类别的样本数量，宏平均和加权平均则提供了模型在各个类别上的整体表现和考虑类别不平衡性的整体表现。从下表可以看出，模型在未患糖尿病和患糖尿病两个类别上的精确率、召回率和 F1 分数都较高，且整体准确率达到了 94%。这表明模型在预测未患糖尿病和患糖尿病时，都有较好的综合表现。

	precision	reacll	f1-score	support
0	0.95	0.95	0.95	347
1	0.93	0.93	0.93	217
accuracy			0.94	564
macro avg	0.94	0.94	0.94	564
weighted avg	0.94	0.94	0.94	564

- **混淆矩阵**



从混淆矩阵中可以看出，XGBoost 分类器在多特征预测患糖尿病的任务中表现良好。模型的准确率接近 94.3%，表明大多数样本被正确分类。假阳性和假阴性数量均为 16，表明模型在预测中有一定的误差，但总体误分类率较低

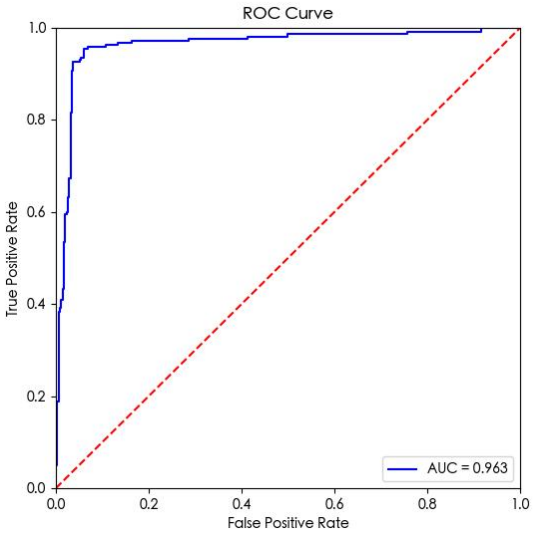
- **ROC 曲线**

ROC 曲线是评估分类模型性能的工具，展示了不同阈值下模型的真阳性率与假阳性率之间的关系。ROC 曲线越接近左上角，模型的性能越好

从下图可以看出：

（1）高 AUC 值：AUC 为 0.963，表明模型在不同阈值下都能保持较高的真阳性率和较低的假阳性率，具有很强的区分能力。

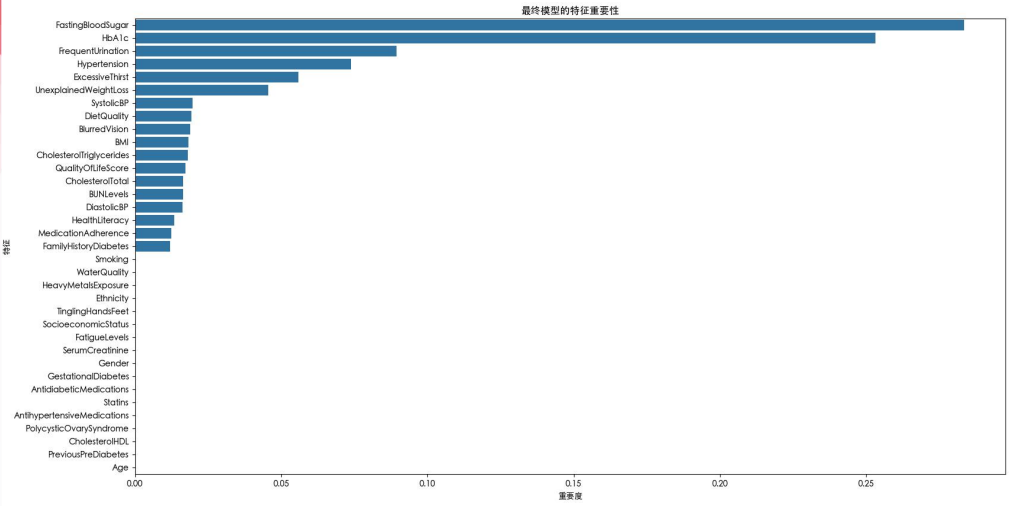
（2）模型性能优异：ROC 曲线接近左上角，进一步验证了 XGBoost 模型在多特征预测患糖尿病任务中的优异表现。



5.4.2. 特征重要性解释

XGBoost 模型的特征重要性程度如下：

特征	重要度
FastingBloodSugar	0.283
HbA1c	0.253
FrequentUrination	0.089
Hypertension	0.074
ExcessiveThirst	0.056
UnexplainedWeightLoss	0.046
SystolicBP	0.020
DietQuality	0.019
BlurredVision	0.019
BMI	0.018
CholesterolTriglycerides	0.018
QualityOfLifeScore	0.017
CholesterolTotal	0.017
BUNLevels	0.016
DiastolicBP	0.016
HealthLiteracy	0.013
MedicationAdherence	0.012
FamilyHistoryDiabetes	0.012



通过模型的特征重要性评估，我们发现与数据探索阶段识别出的关键特征高度一致，这进一步验证了模型的准确性和可靠性。具体来说：

- FastingBloodSugar（空腹血糖）和 HbA1c（糖化血红蛋白）作为糖尿病诊断的核心指标，其重要性在模型预测中得到了充分体现。这两者分别反映了患者的即时血糖水平和长期血糖控制情况，是评估糖尿病病情的关键依据。
- FrequentUrination（尿频）、ExcessiveThirst（过度口渴）和 UnexplainedWeightLoss（不明原因的体重减轻）等临床症状在模型中也占据了重要地位。这些症状的出现往往预示着血糖代谢的异常，是预测糖尿病风险的重要线索。

- Hypertension（高血压）和 DiastolicBP（舒张压）作为心血管疾病的危险因素，与糖尿病的发生和发展紧密相关，高血压是糖尿病的常见并发症。模型将这两个特征纳入考虑，有助于更全面地评估患者的糖尿病风险。
- BMI（身体质量指数）作为肥胖程度的评估指标，在预测糖尿病时同样具有重要意义。肥胖是糖尿病的重要诱因之一，因此通过 BMI 可以初步判断患者的糖尿病风险。
- DietQuality（饮食质量）对于糖尿病的预防和治疗至关重要。模型将饮食质量作为预测特征之一，强调了良好饮食习惯在控制血糖水平、减少并发症方面的重要作用。
- CholesterolTriglycerides（胆固醇和甘油三酯）水平反映了患者的血脂状况，血脂异常是糖尿病及其并发症的重要危险因素。模型将其纳入预测体系，有助于更精确地评估患者的糖尿病风险。
- 虽然 SerumCreatinine（血清肌酐）通常用于评估肾功能，但在某些情况下，如糖尿病肾病等糖尿病并发症中，血清肌酐水平也具有重要的参考价值。模型将其纳入预测特征，体现了对糖尿病并发症风险的全面考虑。

综上所述，这些特征在预测糖尿病时具有较高的重要性是符合医学常识的，侧面证明了我们模型的性能。它们不仅反映了患者的血糖、血脂、血压等生理指标，还体现了患者的临床症状和饮食习惯等方面的信息。通过综合考虑这些特征，我们可以更准确地评估患者患糖尿病的风险，并为其制定个性化的预防和治疗方案。