

**《机器学习》课程论文**

**[](http://bysj.zjgsu.edu.cn/)**

题目：集成学习分类算法在丙型肝炎分类上的应用

**学 院：统计与数学学院**

**专 业：应用统计**

**学 号：22020040146**

**学生姓名：王季昕**

**二○二三 年 四 月**

集成学习分类算法在丙型肝炎分类上的应用

摘要

本文中的数据来自UCI 机器学习数据库，该数据集包含献血者和丙型肝炎患者的实验室值以及年龄等人口统计学值。本文采用GBDT、lightGBM两种集成分类算法分别对丙型肝炎数据集进行进行建模分析，另外选择KNN和SVC两个分类模型作为对比模型，在使用默认模型之后对模型进行调参，通过调参后的集成模型和KNN、SVC模型预测准确率、召回率、F1分数、AUC以及ROC指标的对比，体现出集成学习的效果好于单一模型。对比得出预测效果最好的模型是lightGBM模型，选择该集成模型得到预测结果，提高丙型肝炎预测检验的正确性，具有一定的可行性和临床应用性。

**关键词：**GBDT；lightGBM；KNN；SVM；丙型肝炎

目录

[一、 集成算法 ３](#_Toc312)

[1.1 集成算法简述 ３](#_Toc31042)

[1.2 GBDT介绍 ３](#_Toc16247)

[1.3 lightGBM介绍 ４](#_Toc73)

[二、其他算法 ５](#_Toc26673)

[2.1 SVM介绍 ５](#_Toc17236)

[2.2 KNN介绍 ６](#_Toc901)

[三、模型评价指标 ７](#_Toc28566)

[3.1 混淆矩阵及相关评价指标 ７](#_Toc296)

[3.2 ROC曲线和AUC ７](#_Toc30734)

[四、数据集处理和描述性分析 ９](#_Toc18871)

[4.1数据集介绍 ９](#_Toc17949)

[4.2 数据集分布及数据预处理 １０](#_Toc32634)

[4.2 数据的描述性分析 １０](#_Toc32683)

[4.3 变量相关性分析 １２](#_Toc4718)

[五、集成模型分类 １３](#_Toc6031)

[5.1 GBDT模型预测分析 １３](#_Toc10652)

[5.2 lightGBM模型预测分析 １４](#_Toc16919)

[六、对比分类模型实验 １６](#_Toc13300)

[6.1 SVC模型预测分析 １６](#_Toc30429)

[6.2 KNN模型预测分析 １７](#_Toc17747)

[七、总结 １９](#_Toc12884)

[参考文献 ２０](#_Toc18518)

# 集成算法

## 集成算法简述

集成算法是一种将多个弱分类器或回归器组合成一个强分类器或回归器的方法，可以提高预测的准确性和鲁棒性，从而提高模型的预测能力和泛化能力的机器学习技术。它可以将多个模型的预测结果进行加权或投票的方式进行综合，避免单一模型的过拟合或欠拟合问题，以达到比单个模型更优秀的预测效果。常见的集成算法包括：

1.Bagging（Bootstrap aggregating）：通过随机抽取样本集进行有放回的重复抽样来训练多个模型，然后将它们的预测结果进行投票或平均得到最终的预测结果。常用的Bagging算法包括随机森林。

2.Boosting：通过串行地训练多个弱分类器或回归器，并根据之前分类器的预测误差来调整样本权重，使得下一个分类器能够更关注先前分类错误的样本，从而逐步提升模型性能。常用的Boosting算法包括Adaboost、GBDT、XGBoost、lightGBM等等。

Stacking：通过训练多个不同的分类器或回归器，将它们的预测结果作为新的特征输入到另一个模型中进行训练，从而得到更加准确的预测结果。

集成算法在许多领域都有广泛应用，包括机器学习、数据挖掘和计算机视觉等。本文采用了Boosting策略来集成模型，选择了GBDT、lightGBM两个模型对丙型肝炎数据集进行分类。

## GBDT介绍

GBDT是基于 Boosting 的一种集成学习算法，被广泛用于处理分类，回归问题。该算法限定弱学习器为CART决策树，通过串行训练多个相互依赖的决策树分类模型的方式，采用和算法一样的前向分布式进行多次迭代，每轮迭代产生一个弱分类器来残差，下一轮分类器根据上一个分类器的残差再进行预测，最后将所有预测经过加权处理再相加得出结果。梯度提升决策树对弱分类器的要求一般是足够简单，并且是低方差和高偏差的。因为训练的过程是通过降低偏差来不断提高最终分类器的精度。GBDT模型应用在二分类的场景当中的原理其实和逻辑回归相似，只不过在GBDT的FM是一个加法模型，如下式：



其中，为决策树，是第m棵决策树的参数，为树的数量，是所有决策树结果的和，是第棵决策树的预测结果，是指第m棵决策树。

本文研究问题是分类问题，所以采用交叉熵损失函数来计算参数



所以本文的分类问题转化为求损失函数最小的参数值，如下：



## 1.3 lightGBM介绍

LightGBM 在传统的 GBDT 基础上引入了两个新技术:梯度单边采样(GOSS) 和独立特征合并(EFB)。GOSS 对小梯度样本点进行随机采样,保留对信息增益影响更大的梯度大的样本,在保持信息增益评估的精度前提下,大大提高了模型学习速率,且在采样率相同情况下,梯度单边采样的结果比随机采样准确率更高。 EFR 则实现了互斥特征的捆绑,达到减少特征维度的目的,提高了模型运算效率。 另外,相较于传统 GBDT 算法使用 pre-sorted 算法以精确分割数据,LightGBM 使用了直方图算法,即将连续浮动的特征离散成 k个离散值,并构造宽度为 k 的直方图,大大降低了内存消耗以及数据分割复杂度。

# 二、其他算法

## 2.1 SVM介绍

支持向量机是一种既可以做分类，也可以做回归的算法，若输入标签为连续值则做回归，若输入标签为分类值做分类。它基于结构风险最小化理论，对于线性可分的情况，在特征空间中构建最优分类面，对于线性不可分的情况，通过核函数（线性核函数、高斯核函数、多项式核函数等）将低维输转化为高维特征空间，使其线性可分，从而使得在高维特征空间中采用线性算法对样本的非线性特征进行线性分析成为可能。

[SVM](https://so.csdn.net/so/search?q=SVM&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/SummerStoneS/article/details/_blank)的核心将数据的特征投射到高维，然后找到超平面，分割不同类别的数据点，而且要使分离的程度越大越好，每个类别都会有一些数据点作为支撑向量，这些支撑向量决定了最后分割的超平面。

本文所使用的方法是支持向量机算法中的分类算法SVC，对于多维数据来说，我们需要找到分离超平面，所以首先我们需要定义数据点到分割平面的距离，这里用的是点到直线的距离公式，可以表现为如下形式：



支持向量机的目标是让support vector到分割线的距离的距离最大，分离超平面两侧间隔距离为2r，可表示如下:



所以，目标函数是一个极小极大问题，我们总可以通过缩放w和b，使该不等式右边划归到1。如此，就可以简化目标函数。考虑到凸函数的性质，将原目标函数转化成对偶形式，因为目标函数是二次的，而约束在参数w和b上是线性的，因此这个问题是一个凸优化问题，可以通过标准的拉格朗日乘子来求解，对偶的目标函数如下：



## 2.2 KNN介绍

KNN（k-Nearest Neighbors）算法是一种基本的分类和回归算法，它的基本思想是找到最近的k个邻居，并根据这些邻居的标签进行分类或回归。KNN算法在分类问题中经常使用，并且被认为是最简单的机器学习算法之一。

KNN算法的优点在于简单易懂，对于非线性的数据具有很好的分类效果。但是，它的缺点也比较明显，主要包括：

(1)对异常值敏感：如果训练集中存在异常值，可能会影响分类结果。

(2)精度问题：如果k值过小，可能会受到噪声的影响，如果k值过大，可能会受到样本分布的影响。

总之，KNN算法是一种简单有效的算法，但在实际应用中需要根据具体情况进行优化和调整。

# 三、模型评价指标

## 3.1 混淆矩阵及相关评价指标

本题所建立的模型为二分类模型，因此为更好的评价模型，这里采取混淆矩阵来计算模型的准确率、精确率、召回率、F1值、。二分类混淆矩阵如下所示

表1 评价指标表

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 真实标签 | |
| 1 | 0 |
| 预测标签 | 1 | TP | FP |
| 0 | FN | TN |

通过混淆矩阵所计算得到的评价指标分别为：准确率、精确率、召回率以及F1值。

准确率是指全部正确预测数量的比率。公式为：



精确率是指正确预测为某一类的比率。公式为：

，

召回率是指某一类中被正确预测的比率。公式为：



F1值是精确率与召回率的调和平均数。公式为：

。

## 3.2 ROC曲线和AUC

ROC（Receiver Operating Characteristic）曲线是一种用于评估分类模型性能的方法。它通过绘制真阳性率（True Positive Rate，TPR）与假阳性率（False Positive Rate，FPR）的曲线来显示模型的性能。

在ROC曲线上，横轴代表FPR，纵轴代表TPR。对于一个分类模型，预测结果可以被归类为真阳性（True Positive，TP）、假阳性（False Positive，FP）、真阴性（True Negative，TN）、假阴性（False Negative，FN）四种情况。

其中:

ROC曲线的形状描述了分类器的性能，通常ROC曲线越靠近左上角，分类器的性能越好。此外，ROC曲线下面积（AUC）也是评估分类器性能的常用指标。AUC的取值范围在0到1之间，1表示完美分类器，0.5表示随机分类器。ROC曲线经常被用来比较不同分类器的性能，以及寻找最佳的分类器阈值。

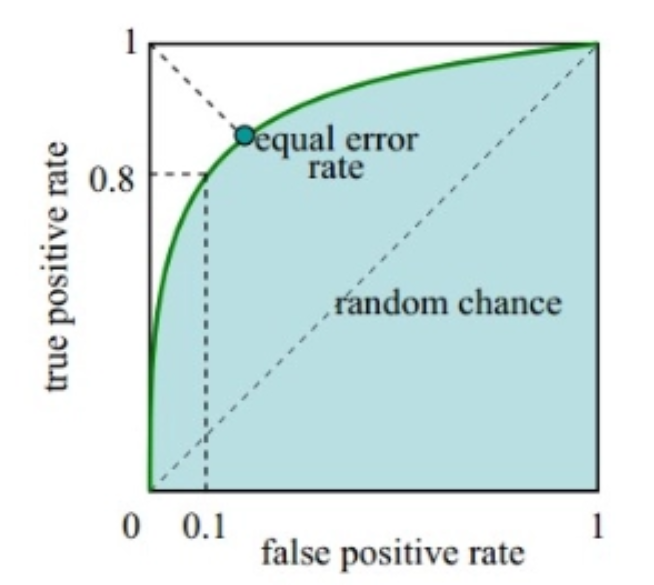


图 1 ROC曲线和AUC值

AUC（Area Under the Curve）是指ROC曲线下面积,取值范围在0到1之间，是一种用于评估分类模型性能的指标，越接近1表示模型性能越好，越接近0.5表示模型性能越差（等同于随机猜测的效果），而取值小于0.5则说明分类器的预测结果比随机猜测还要差。AUC是一个衡量分类器在所有可能阈值下性能的综合指标。与其他单一指标（如准确率、召回率、F1-score等）相比，AUC更能全面评估分类器的性能，并且相对不受类别分布不均的影响。在实际应用中，AUC常被用来比较不同分类器的性能，以及选择最佳分类器阈值。

# 四、数据集处理和描述性分析

## 4.1数据集介绍

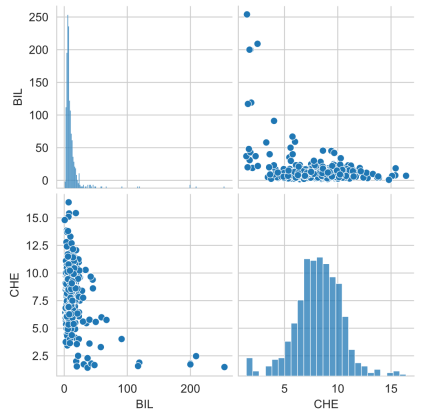
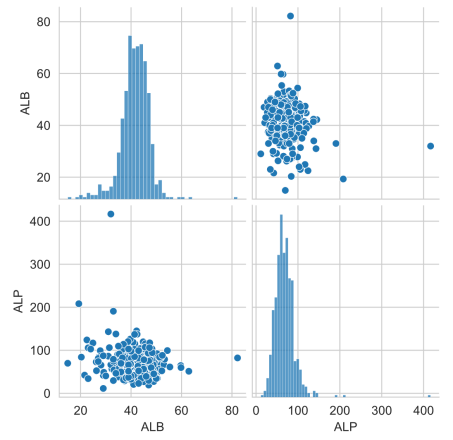
|  |  |  |  |
| --- | --- | --- | --- |
| Type | Field Name | Field Description | Normal value range |
| Continuous feature | Age | - | F and M |
| ALB | albumin | 35-55g/L |
| ALP | alkaline phosphatase | 0-40U/L |
| ALT | Glutamicpyruvic transaminase | M:5-40U/L  F:5-35U/L |
| AST | Glutamic oxaloacetic transaminase | 0-40U/L |
| BIL | bilirubin | 5.10-19umoi/L |
| CHE | Serum cholinesterase | 4.3-10.5U/L |
| CHOL | total cholesterol | 2.83-5.18mmol/L |
| CREA | Creatinine substance | M:50-110umol/L  F:40-100umol/L |
| GGT | Glutamyl transpeptidase | 3-50U/L |
| PROT | Total protein | 20-80mg/L |
| Classification  feature | Sex | - | F and M |

本文中的数据来自UCI 机器学习数据库，由汉诺威医科大学临床化学研究所、赫姆霍兹感染研究中心等研究机构贡献。该数据集包含献血者和丙型肝炎患者的实验室值以及年龄等人口统计学值。主要特征是肝功能和各种成分在血液中的含量，通过对数据集的观察，我们知道数据集主要是数值特征，分类特征只有性别。这些字段的具体含义

如表所示:

## 4.2 数据集分布及数据预处理

首先对丙型肝炎数据数据集进行清洗，数据集不存在缺失值超过5%的特征，利用原则进行数据清洗，发现有三个离群值点，将这些个体剔除。通过下图可以查看部分连续型变量的分布情况，可以看出CREA、BIL等几个变量存在尖峰分布，数据分布非常集中。



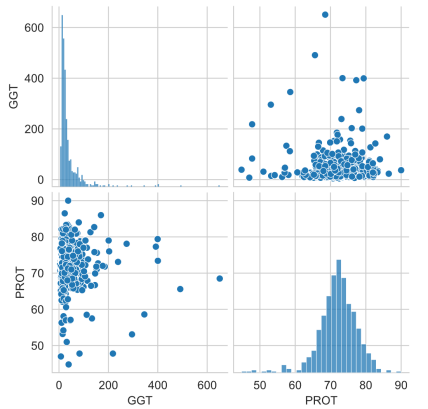
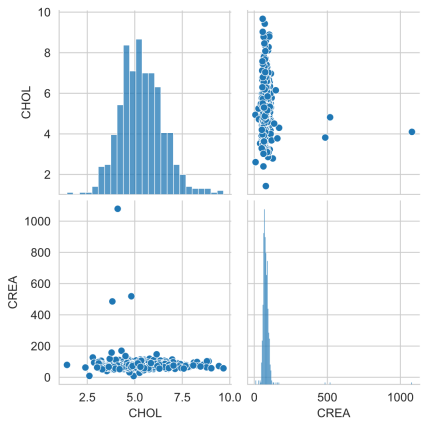


图 2 数据分布

## 4.2 数据的描述性分析

首先对样本标签进行可视化，如左图所示，丙型肝炎的患者样本较少仅有75条，未患病的的样本有545条，样本比例不均衡，所以在下面进行模型训练的时候为了提高模型的泛化能力进行SMOTE过采样方法的方式进行训练，SMOTE法过采样即合成少数类过采样技术，它是一种针对随机过采样容易导致过拟合问题的改进方案，步骤如下：

1、对于少数类中每一个样本x，以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其k近邻。

2、根据样本不平衡比例设置一个采样比例以确定采样倍率N，对于每一个少数类样本x，从其k近邻中随机选择若干个样本，从而构造新的样本，采用SMOTE过采样方法得到新的训练样本的正负样本个数如右图所示：

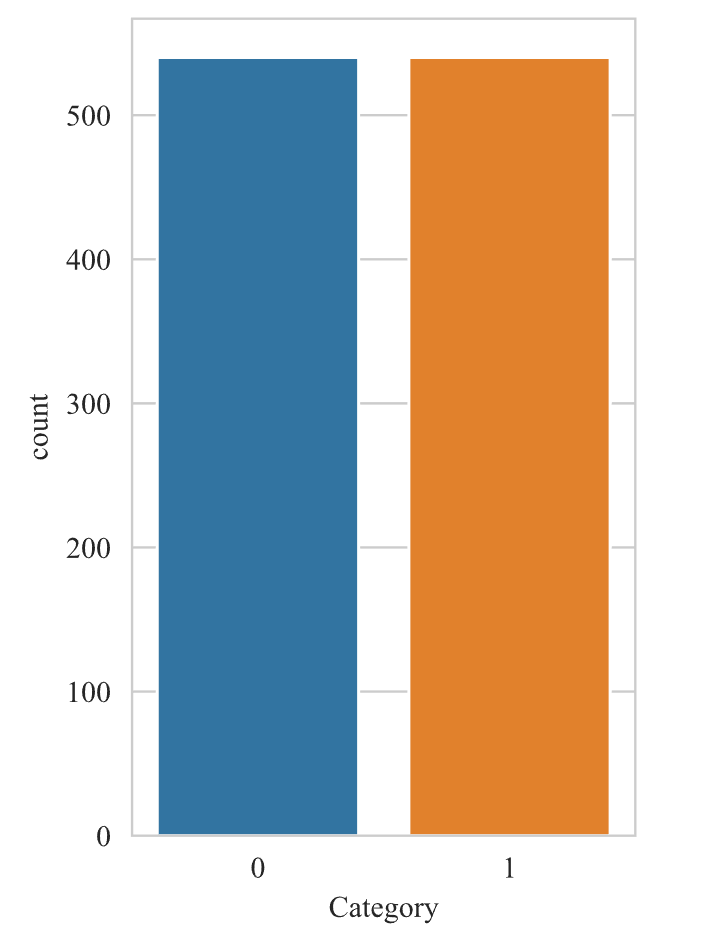
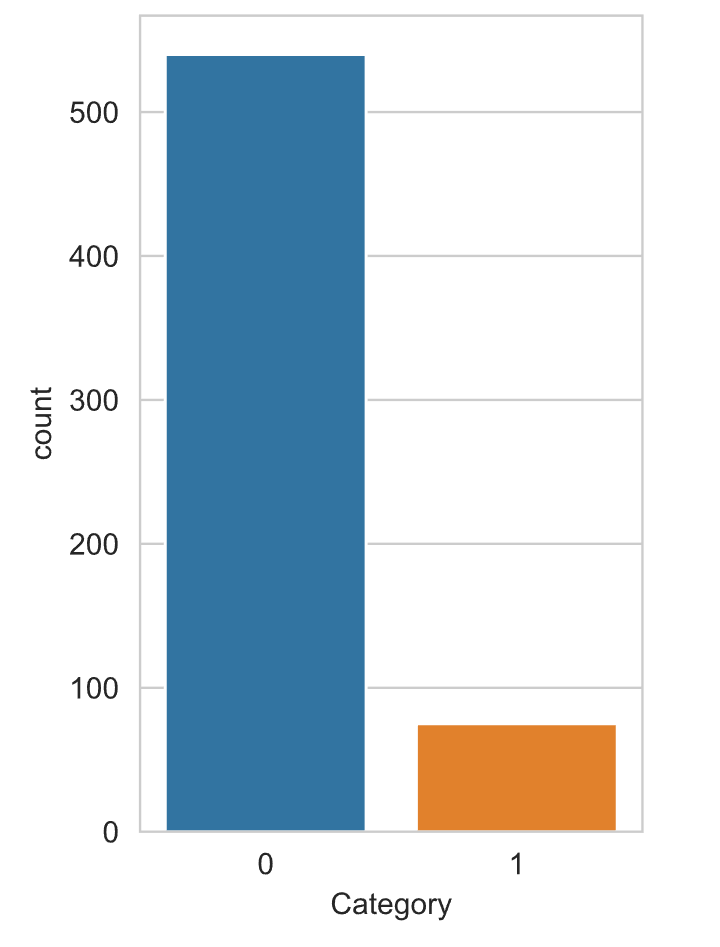


图 3 过采样前后样本标签分布对比

下图是探究性别对患病率的影响，由下图可见，不同性别的患病率大致相同，说明对于不同性别人群，丙型肝炎的患病率大致形同。

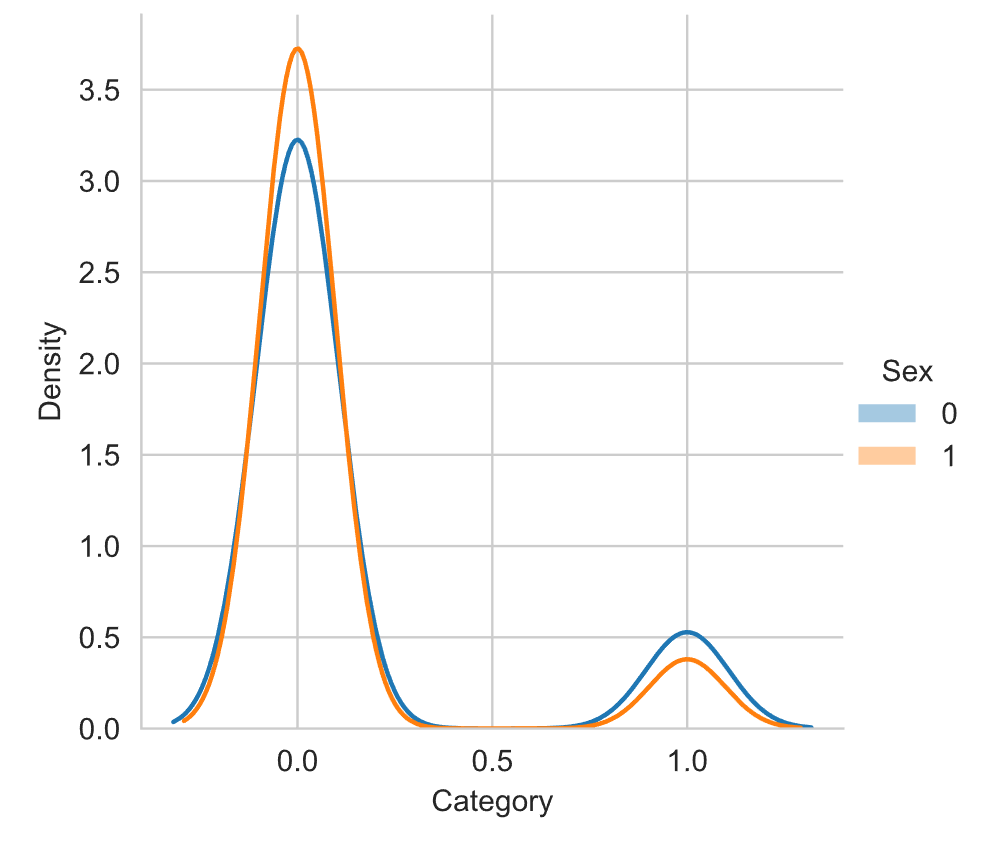


图 4 不同性别人群丙型肝炎患病率分布

下图是不同年龄段的发病率的分布图，由图可以看出，在该数据中的大部分数据都集中在30-50岁之间。在不同年龄段的丙型肝炎患病率不同，分布略有偏差。年龄越大丙型肝炎患病率会提高。

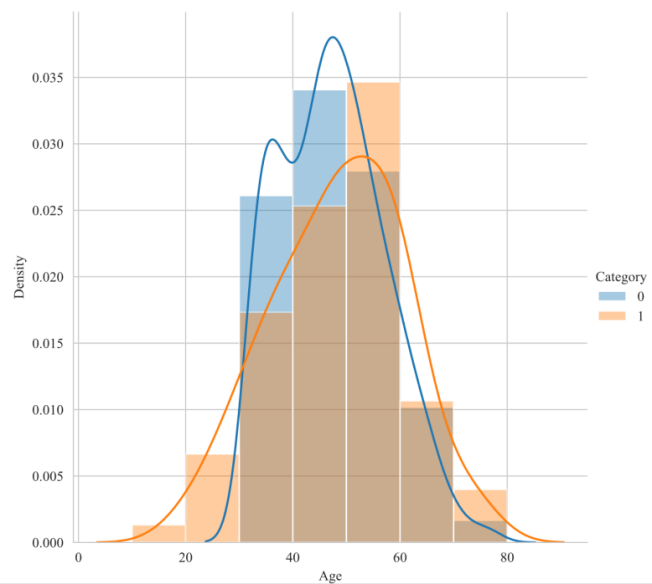


图 5 不同年龄段丙型肝炎患病率分布

## 4.3 变量相关性分析

由如下变量相关性热力图可以看出自变量之间不存在强相关变量，故保留所有相关变量。对分类结果影响最大的指标就是AST、GGT和BIL三个指标，对分类及结构影响最小的指标是Age、ALP、Sex、PROT四个指标。

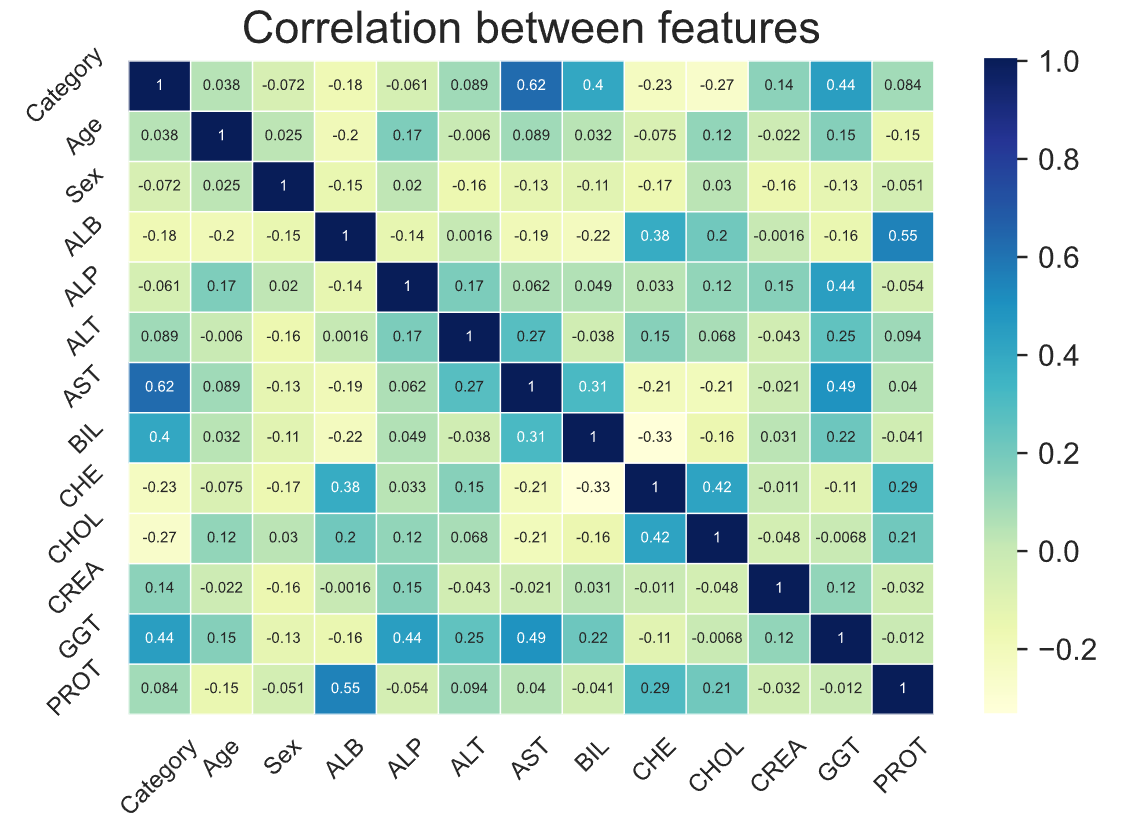


图 6 变量相关性热力图

# 五、集成模型分类

## 5.1 GBDT模型预测分析

左图是使用默认GBDT模型用来进行丙型肝炎数据集分类的模型ROC曲线和AUC值，可以看出使用默认的GBDT的模型效果就很好，准确度可以达到98%，右图是经过调参之后的模型ROC曲线和AUC值，可以看出调参之后的ROC曲线基本没有变动，AUC值略有提升。由于初始模型的预测效果太好，调参后模型基本没有提升效果。

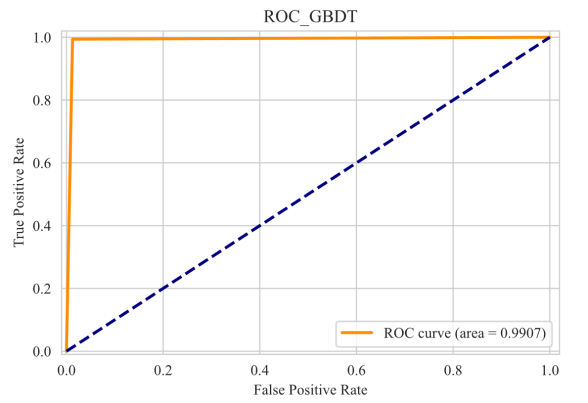
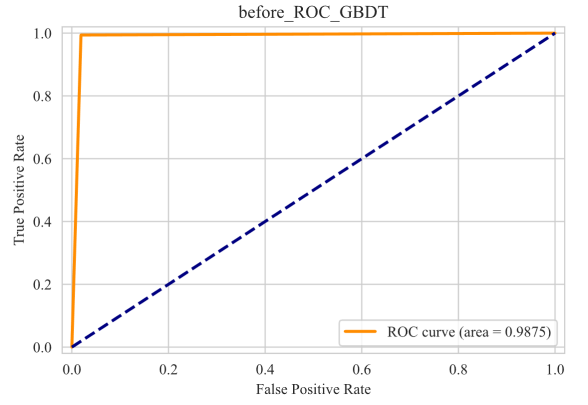


图 7 GBDT模型调参前后的ROC曲线和AUC值

参数优化：

由于GBDT的训练速度较慢，而且默认GBDT模型的训练效果很好，故参数只调整n\_estimators。

1. n\_estimators: 默认是100。弱学习器的个数。一般来说n\_estimators太小，容易欠拟合，n\_estimators太大，又容易过拟合，一般选择一个适中的数值。

调参前后的各指标对比：

表2 GBDT参数优化前评级指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 准确率 | F1值 | 精准率 | 召回率 | AUC |
| 默认参数 | 0.9932 | 0.9908 | 0.9923 | 0.9878 | 0.9875 |
| n\_estimators:450 | 0.9907 | 0.9908 | 0.9939 | 0.987878 | 0.9907 |

下图是调参后模型的混淆矩阵，可以看出分类效果很好，基本没有错误分类。

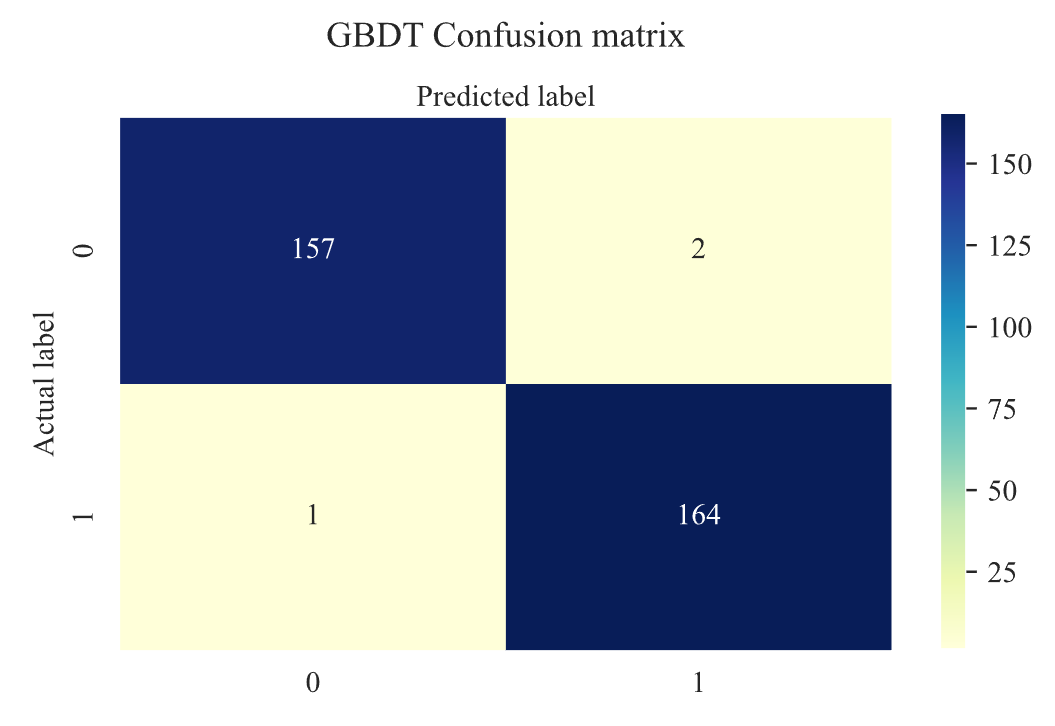


图 8 GBDT混淆矩阵

## 5.2 lightGBM模型预测分析

LightGBM相比较于GBDT的效果更好，左图是使用默认lightGBM模型用来进行丙型肝炎数据集分类的模型ROC曲线和AUC值，可以看出使用默认的lightGBM的模型效果就很好，准确度可以达到99%，右图是经过调参之后的模型ROC曲线和AUC值，可以看出调参之后的ROC曲线基本没有变动，AUC值还略有下降。由于初始模型的预测效果太好，调参后模型基本没有提升效果。

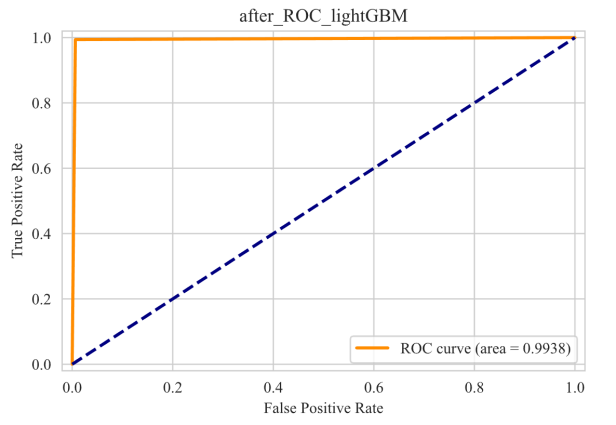
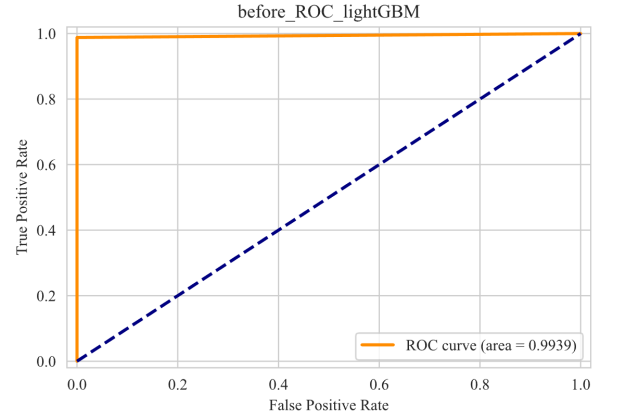


图 9 lightGBM模型调参前后的ROC曲线和AUC值

调参优化：

1. max\_depth：default默认不限制深度，该参数用于限制树模型的最大深度。这可以在数据集小的情况下防止过拟合，max\_depth的选择需要考虑到数据量、特征维度和连续特征的数量。
2. n\_estimators：默认值为100，表示树的数量。通常在100-1000之间进行调整。增加树的数量可以提高模型的性能，但也会增加模型的计算时间。需要根据数据量和模型复杂度进行适当调整，以达到平衡性能和效率的目的。

调参前后的各指标对比：

表3 lightGBM参数优化前评级指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 准确率 | F1值 | 精准率 | 召回率 | AUC |
| 默认参数 | 0.9938 | 0.9939 | 1.0 | 0.9878 | 0.9939 |
| **max\_depth：3**  **n\_estimators：340** | 0.9969 | 0.9969 | 0.9938 | 1.0 | 0.9938 |

下图是调参后模型的混淆矩阵，可以看出分类效果很好，基本没有错误分类。

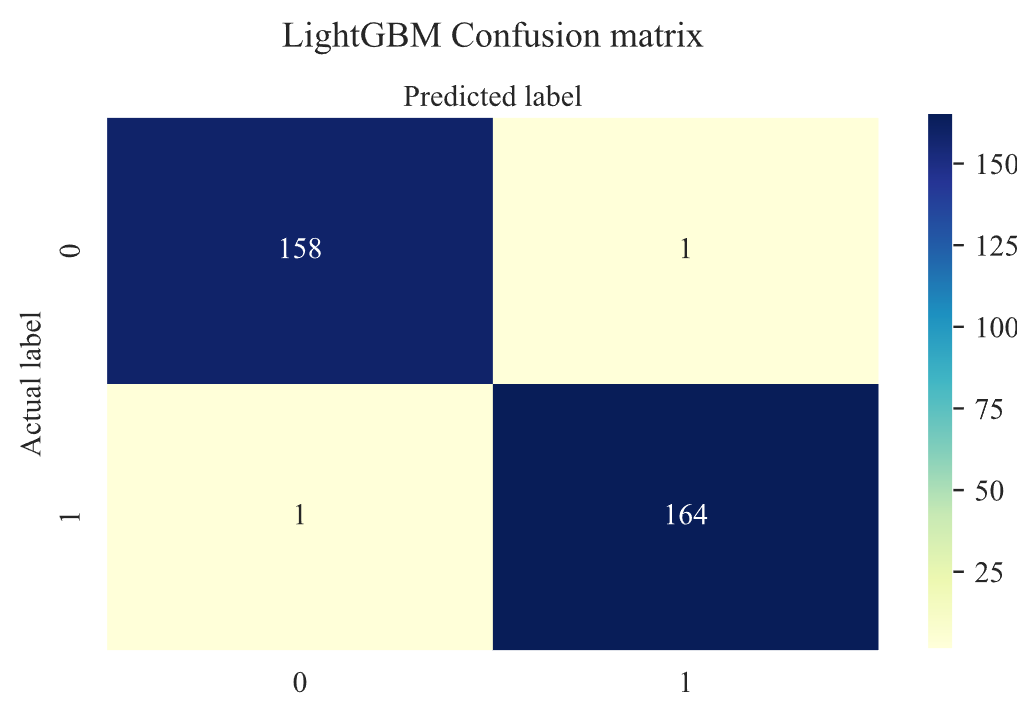


图 10 lightGBM混淆矩阵

# 六、对比分类模型实验

## 6.1 SVM模型预测分析

左图是使用默认SVM模型，和函数选择sigmoid核函数，用来进行丙型肝炎数据集分类的模型ROC曲线和AUC值，右图是经过调参之后的模型ROC曲线和AUC值，可以看出调参之后，的ROC曲线明显向左上角靠近和AUC值增大大概7%，模型提升效果显著。

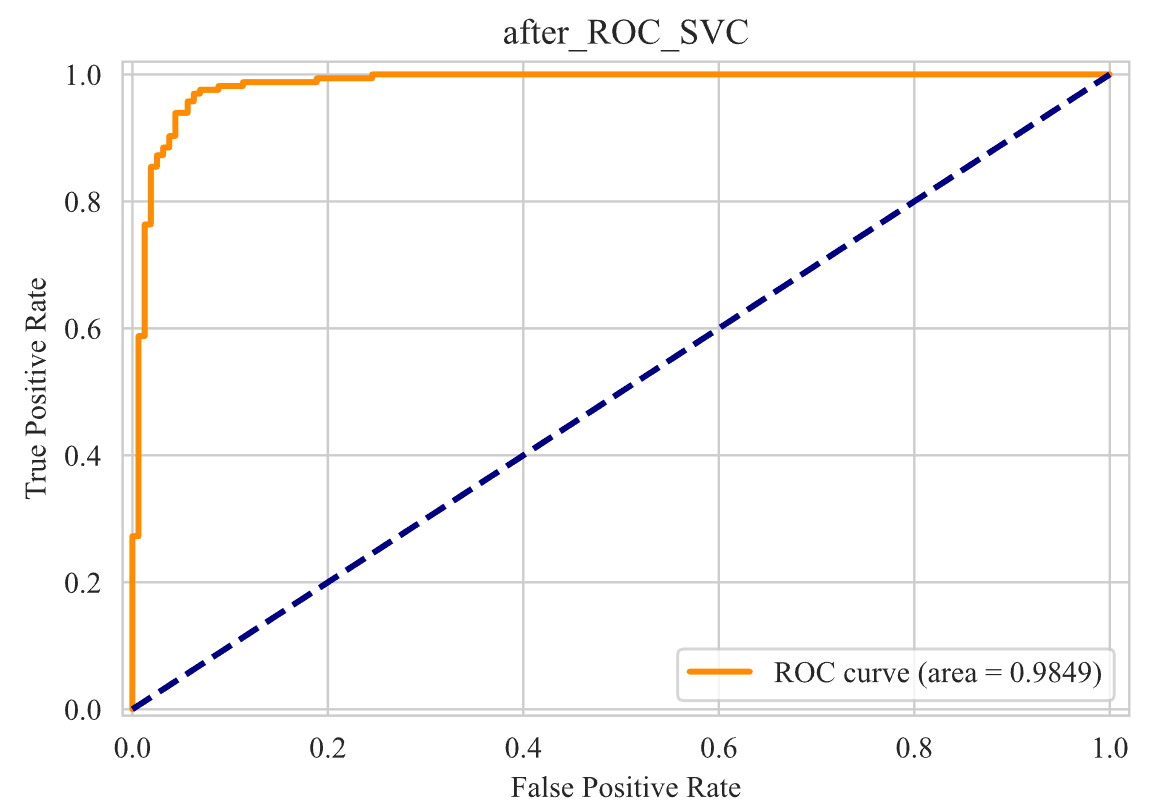
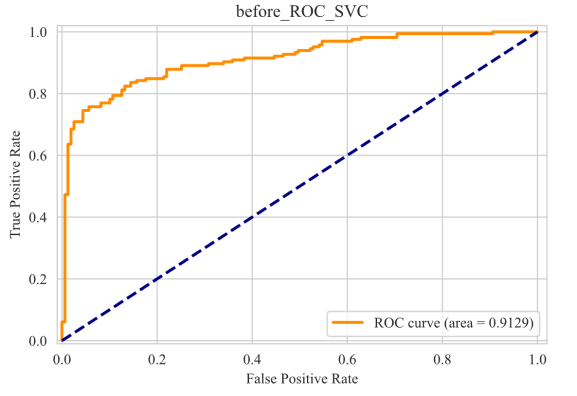


图 11 svc模型调参前后模型ROC曲线和AUC值

SVM调参：

（1）C：惩罚参数，C越大，对训练集测试时准确率很高，但泛化能力弱。C值小，对误分类的惩罚减小，允许容错，将他们当成噪声点，泛化能力较强。

（2）Kernel：核函数用于将数据从低维映射到高维,以便找到更好的分类超平面的函数，常见核函数：linear, poly, rbf, sigmoid。

调参前后的各指标对比：

表4 SVM参数优化前评级指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 准确率 | F1值 | 精准率 | 召回率 | AUC |
| 默认参数 | 0.7746 | 0.7623 | 0.9791 | 0.5969 | 0.9129 |
| C:1,kernel=sigmoid | 0.9351 | 0.9337 | 0.9736 | 0.8869 | 0.9800 |

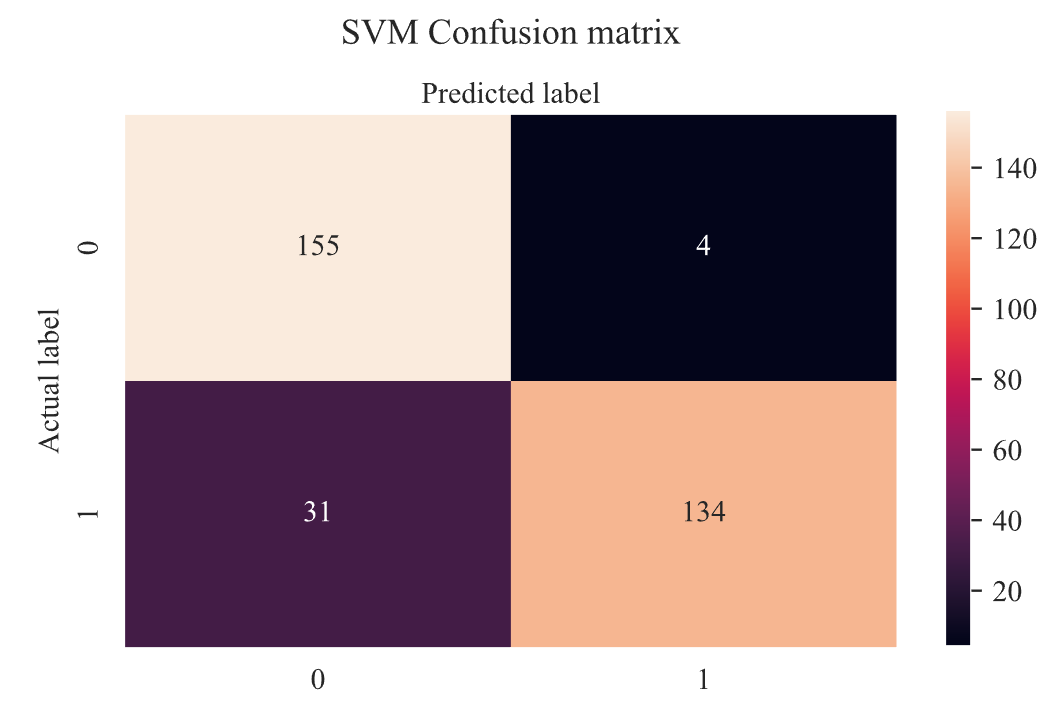


图 12 svm混淆矩阵

## 6.2 KNN模型预测分析

左图是使用默认KNN模型用来进行丙型肝炎数据集分类的模型ROC曲线和AUC值，右图是经过调参之后的模型ROC曲线和AUC值，可以看出调参之后，的ROC曲线明显向左上角靠近和AUC值增大大概2%，模型提升效果一般。

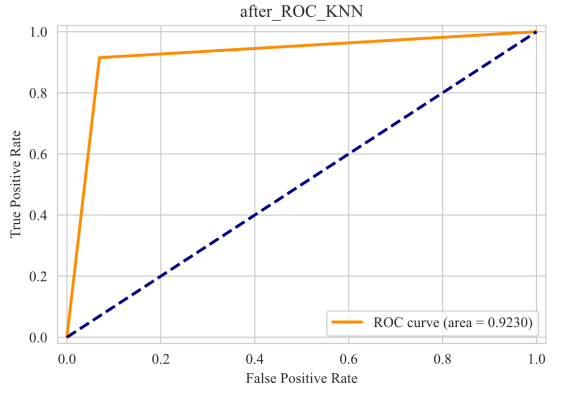
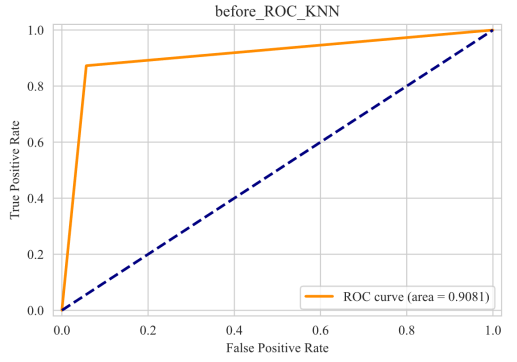


图 13 KNN模型调参前后ROC曲线和AUC值

KNN调参：

1. K:在给定一个训练集中，输入一个新数据，然后找寻新的输入实例在训练集中最邻近的k个实例。这个k个实例的多数属于某个类，就把该输入实例分类到这个类中.

下表是KNN调参前后的各模型评价指标对比图，可以看出调参前后个指标有明显提升：

表5 KNN参数优化前评级指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 准确率 | F1值 | 精准率 | 召回率 | AUC |
| 默认参数 | 0.9012 | 0.9018 | 0.9130 | 0.8909 | 0.9081 |
| n\_neighbors = 3 | 0.9960 | 0.9942 | 1.0 | 0.9942 | 0.9230 |

下图是混淆矩阵的可视化，可以看出只有少部分样本被分类错误：

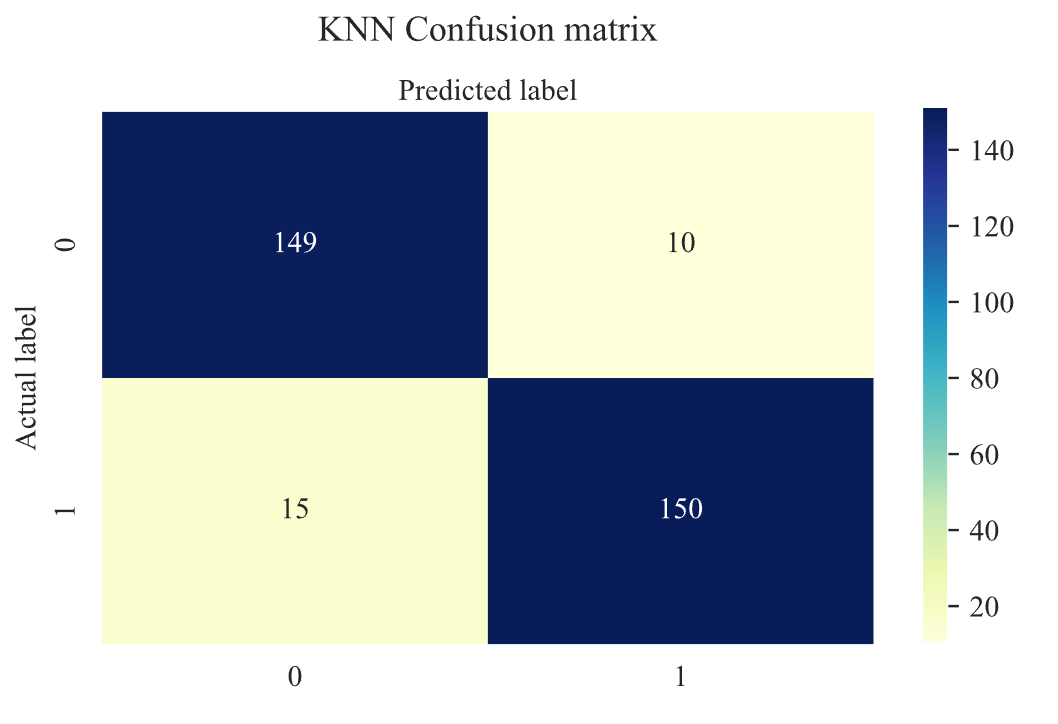


图 14 KNN混淆矩阵

# 七、总结

为了更直观地对比集成学习GBDT、lightGBM和其他分类算法SVM和KNN的表现，将四种算法经过调参之后的模型在测试集上的准确率、精确率、召回率、F1得分对比如下：

表 6 四种分类模型的各指标对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 算法 | 准确度 | 精确率 | 召回率 | F1得分 | AUC |
| SVM | 0.9351 | 0.9736 | 0.8969 | 0.9337 | 0.9800 |
| KNN | 0.9960 | 1.0 | 0.9929 | 0.9942 | 0.9230 |
| GBDT | 0.9907 | 0.9939 | 0.9878 | 0.9908 | 0.9907 |
| LightGBM | 0.9969 | 1.0 | 0.9938 | 0.9969 | 0.9938 |

可见对于一般的单模型分类算法，默认模型效果一般，但是经过调参以后，模型的准确度和精确率、召回率、F1得分等指标均会有所上升。对于集成算法来说，由于集成了多个分类模型，初始模型的的分类准确率就会很高，对于本数据集来说，参数调整对模型效果的提升有限，基本经过调参之后的模型准确率能达到99%以上。对于本文采取的两种集成分类模型GBDT和lightGBM，初始模型的效果就很好，展现了集成模型的强大效果，调参后模型效果提升不大。两个模型的表现相近，但是lightGBM的模型训练速度更快，所以选择lightGBM模型作为最后的预测模型。

参考文献

[1] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information pROCessing systems, 2017, 30.

[2] Chen C, Zhang Q, Ma Q, et al. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion[J]. Chemometrics and Intelligent Laboratory Systems, 2019, 191: 54-64.

[3] Wang D, Zhang Y, Zhao Y. LightGBM: an effective miRNA classification method in breast cancer patients[C]//PROCeedings of the 2017 international conference on computational biology and bioinformatics. 2017: 7-11.

[4] Zhang W, Yu J, Zhao A, et al. Predictive model of cooling load for ice storage air-conditioning system by using GBDT[J]. Energy Reports, 2021, 7: 1588-1597.

[5] 曾梓铭. 基于GBDT的医药供应链金融信用风险评价研究[D].桂林电子科技大学,2022.DOI:10.27049/d.cnki.ggldc.2022.000997.

[6] Guo G, Wang H, Bell D, et al. KNN model-based approach in classification[C]//On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. PROCeedings. Springer Berlin Heidelberg, 2003: 986-996.

1. Wang H, Hu D. Comparison of SVM and LS-SVM for regression[C]//2005 International conference on neural networks and brain. IEEE, 2005, 1: 279-283.