

数据挖掘

题目：基于Kaggle心脏病数据集的分类算法研究

学 院 计算机与电子信息学院

专 业 计算机科学与技术

班 级 计科161

姓 名 韩鹏宇

学 号 1607300314

指导老师 兰伟

2019 年 1 月 2日

目录

[一．绪论 2](#_Toc28983368)

[1 实验背景 2](#_Toc28983369)

[二．算法概述 3](#_Toc28983370)

[1.逻辑回归算法： 3](#_Toc28983371)

[2.决策树算法： 4](#_Toc28983372)

[2.1三种决策树算法： 4](#_Toc28983373)

[3. kNN算法：K最近邻(kNN，k-NearestNeighbor)分类算法 6](#_Toc28983374)

[3.实验数据及结果 7](#_Toc28983375)

[3.1数据集选择及预处理 7](#_Toc28983376)

[3.2 软件实现 9](#_Toc28983377)

[3.2.1 逻辑回归算法 10](#_Toc28983378)

[3.2.2 kNN 算法： 11](#_Toc28983379)

[3.2.3 决策树 12](#_Toc28983380)

[3.3结果讨论 14](#_Toc28983381)

[4.总结与展望 15](#_Toc28983382)

# 一．绪论

## 1 实验背景

心脏病作为一种心血管疾病在生活中随处可见。目前，心血管疾病是我国威胁人民身体健康的几类疾病之一，这种疾病已经严影响到我国经济发展的重大问题。面对如此严峻的形势，如何对心脏病的诊断和治疗进行干预已经成为迫在眉睫需要解决的问题。而在信息化迅速发展的时代，将电子科学技术与医学的心脏疾病相结合，是现在研究人员的热门研究问题。经过近几年医学研究人员的不断努力，对心脏病的诊断研究已初见成效，也积累了大量数据供研究人员参考。近几年在世界上迅速发展的交叉学科是数据挖掘学科，他涉及的范围相对比较广泛，包括有统计学学科，人工智能学科，还有机器学习学科，还涉及数据库学科等多个领域。随着科学技术的发展，数据库管理系统应用的范围也相对广泛，对于患者的人数积累也越来越多，数据库信息的增长速度十分快，犹如汪洋大海。若没有强大的处理工具，对数据信息的理解已经超出研究人员的理解能力。因此在数据库中的大型数据由于不能被理解，几乎变成了永埋地下的人们很难再从中提取数据档案，而越来越多的患者数据中隐含着许多重要的信息，决策者常常不是因为数据库中的数据去决定一项决策，而是凭自己的主观臆断。因为他们对提取数据库中有用信息的认识是不足的。

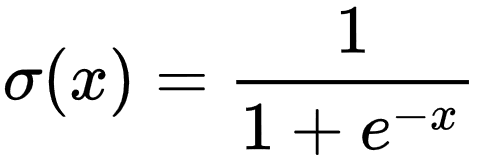
心脏病的诱发病因一直是有很多种情况的，由于有多种情况导致心脏病，因此，根据Kaggle上提供的心脏病数据集中，与心脏病有关的特征有年龄、性别、胸痛类型、静息血压、血清胆固醇、空腹血糖、静息心电图、最大心率、运动诱

发、血管情况、是否生病等，进行分类算法实验，对比分析各种算法的区别。

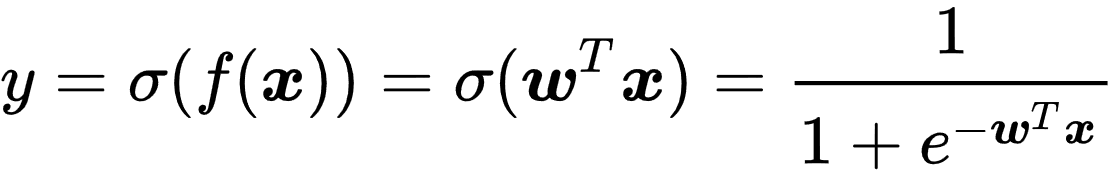
# 二．算法概述

## 1.逻辑回归算法：

logistic回归又称logistic回归分析，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此因变量就为是否胃癌，值为“是”或“否”，自变量就可以包括很多了，如年龄、性别、饮食习惯、幽门螺杆菌感染等。自变量既可以是连续的，也可以是分类的。然后通过logistic回归分析，可以得到自变量的权重，从而可以大致了解到底哪些因素是胃癌的危险因素。同时根据该权值可以根据危险因素预测一个人患癌症的可能性。

逻辑回归公式：（1）sigmoid函数：

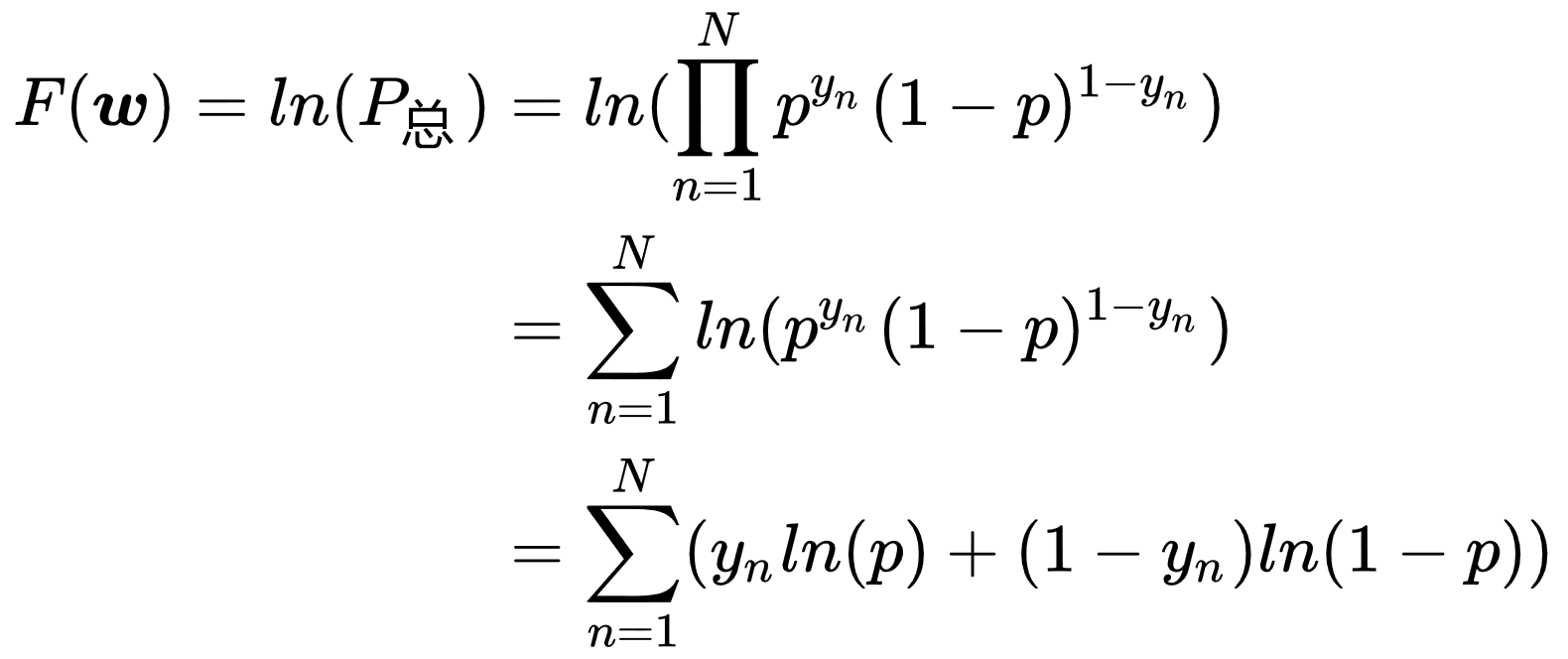
（2）结合sigmiod函数，变成逻辑回归模型：

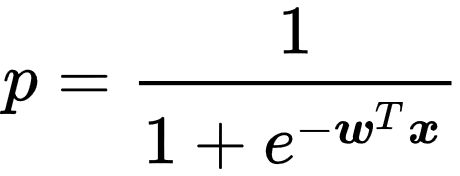


（3）逻辑回归的损失函数（Loss Function）：

损失函数就是用来衡量模型的输出与真实输出的差别。

假设只有两个标签1和0。我们把采集到的任何一组样本看做一个事件的话，那么这个事件发生的概率假设为p。我们的模型y的值等于标签为1的概率也就是p。





这个函数F(w)又叫做它的损失函数。损失函数可以理解成衡量我们当前的模型的输出结果，跟实际的输出结果之间的差距的一种函数。这里的损失函数的值等于事件发生的总概率，我们希望它越大越好。但是跟损失的含义有点儿违背，因此也可以在前面取个负号。

## 2.决策树算法：

决策树(Decision Tree）是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。Entropy = 系统的凌乱程度，使用算法ID3, C4.5和C5.0生成树算法使用熵。这一度量是基于信息学理论中熵的概念。

决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

分类树（决策树）是一种十分常用的分类方法。他是一种监管学习，所谓监管学习就是给定一堆样本，每个样本都有一组属性和一个类别，这些类别是事先确定的，那么通过学习得到一个分类器，这个分类器能够对新出现的对象给出正确的分类。这样的机器学习就被称之为监督学习。

### 2.1三种决策树算法：

**1. ID3算法**

ID3算法的核心是在决策树各个结点上对应信息增益准则选择特征，递归地构建决策树。

具体方法是：

1）从根结点(root node)开始，对结点计算所有可能的特征的信息增益，选择信息增益最大的特征作为结点的特征。

2）由该特征的不同取值建立子节点，再对子结点递归地调用以上方法，构建决策树；直到所有特征的信息增益均很小或没有特征可以选择为止；

3）最后得到一个决策树。

ID3相当于用极大似然法进行概率模型的选择

**2. C4.5的生成算法**

与ID3算法相似，但是做了改进，将信息增益比作为选择特征的标准。

**3. CART算法**

CART算法是一种二分递归分割技术，把当前样本划分为两个子样本，使得生成的每个非叶子结点都有两个分支，

   因此CART算法生成的决策树是结构简洁的二叉树。由于CART算法构成的是一个二叉树，它在每一步的决策时只能

   是“是”或者“否”，即使一个feature有多个取值，也是把数据分为两部分。在CART算法中主要分为两个步骤。

CART生成算法（分类树）

在这里需要提一下基尼系数：

在分类问题中，假设有KK类，样本点属于第kk类的概率为pkpk,则概率分布的基尼指数定义为：

Gini(p)=∑Kk=1pk(1−pk)=(p1+p2+...+pK)−∑Kk=1p2k=1−∑Kk=1p2kGini(p)=∑k=1Kpk(1−pk)=(p1+p2+...+pK)−∑k=1Kpk2=1−∑k=1Kpk2

对于分类问题：设CkCk为DD中属于第kk类的样本子集，则基尼指数为：

Gini(D)=1−∑Kk=1(|Ck||D|)2Gini(D)=1−∑k=1K(|Ck||D|)2

设条件AA将样本DD切分为D1D1和D2D2两个数据子集，则在条件AA下的样本DD的基尼指数为：

Gini(D,A)=|D1|DGini(D1)+|D2|DGini(D2)Gini(D,A)=|D1|DGini(D1)+|D2|DGini(D2)

注意：基尼指数也表示样本的不确定性，基尼指数值越大，样本集合的不确定性越大。

算法实现步骤：

1）计算现有样本DD的基尼指数，之后利用样本中每一个特征AA，及AA的每一个可能取值aa，根据A>=aA>=a与A<aA<a将样本分为两部分，并计算Gini(D,A)Gini(D,A)值

2）找出对应基尼指数最小Gini(D,A)Gini(D,A)的最优切分特征及取值，并判断是否切分停止条件，否，则输出最优切分点

3）递归调用1）2）

4）生成CART决策树

## 3. kNN算法：K最近邻(kNN，k-NearestNeighbor)分类算法

邻近算法，或者说K最近邻(kNN，k-NearestNeighbor)分类算法是数据挖掘分类技术中最简单的方法之一。所谓K最近邻，就是k个最近的邻居的意思，说的是每个样本都可以用它最接近的k个邻居来代表。

kNN算法的核心思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。 kNN方法在类别决策时，只与极少量的相邻样本有关。由于kNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，kNN方法较其他方法更为适合。

# 3.实验数据及结果

## 3.1数据集选择及预处理

1.数据来源于Kaggle数据集

（1）数据的基本信息如下表3.1.1

|  |  |
| --- | --- |
| age 年龄 | 303 non-null int64 |
| sex 性别 | 1=male,0=female |
| cp 胸痛类型(4种) | 值1:典型心绞痛，值2:非典型心绞痛，值3:非心绞痛，值4:无症状 |
| trestbps 静息血压 | 303 non-null int64 |
| chol 血清胆固醇 | 303 non-null int64 |
| fbs 空腹血糖 | >120mg/dl ,1=true; 0=false |
| restecg 静息心电图 | 0,1,2 |
| thalach 达到的最大心率 | 303 non-null int64 |
| exang 运动诱发的心绞痛 | 1=yes;0=no |
| oldpeak 相对于休息的运动引起的ST值 |  |
| slope 运动高峰ST段的坡度 | Value 1: upsloping向上倾斜, Value 2: flat持平, Value 3: downsloping向下倾斜  ca The number of major vessels(血管) (0-3) |
| thal A blood disorder called thalassemia | (3 = normal; 6 = fixed defect; 7 = reversable defect)  一种叫做地中海贫血的血液疾病(3 =正常;6 =固定缺陷;7 =可逆转缺陷 |
| target 生病没有 | (0=no,1=yes) |

表3.1.1 数据集特征含义

（2）查看特征之间的一些关系，如图3.1.2、3.1.3、3.1.4

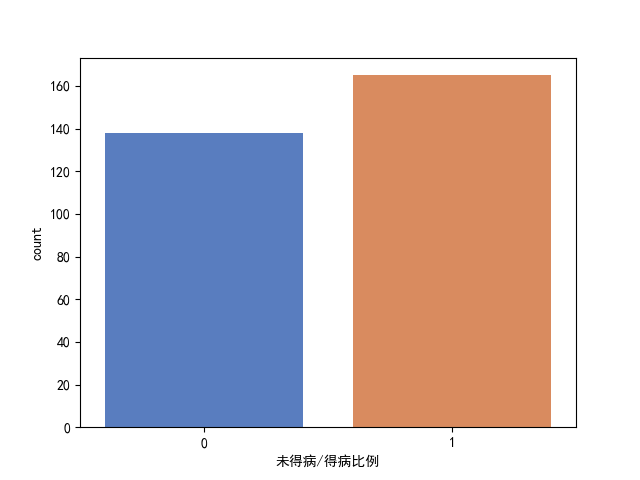


图3.1.2 得病比例

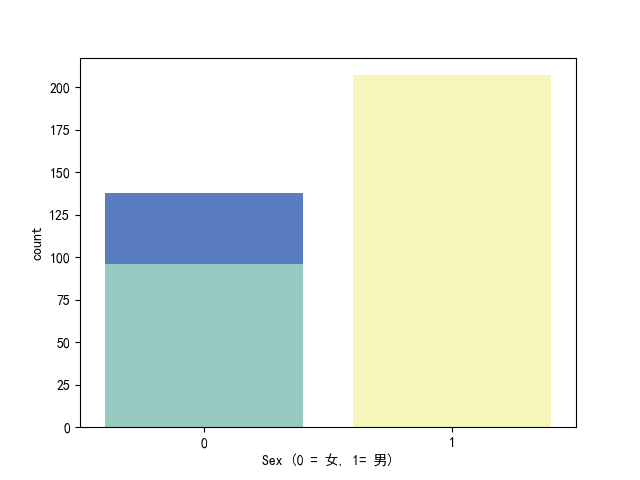


图3.1.3 特征中男女比例

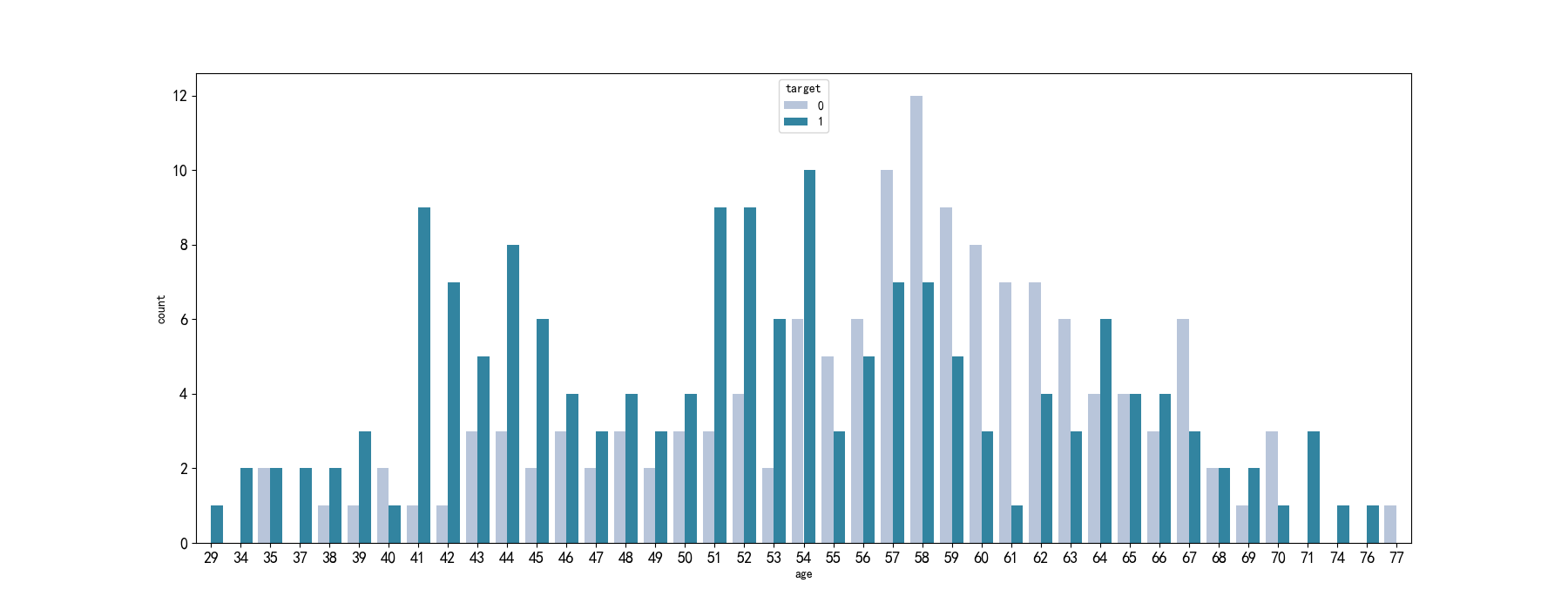


图3.1.4总特征关系

（3）数据处理

特征中有非连续型数值（cp,slope,thal）特征进行处理

处理后的数据为 22\*8 shape为（303,21）

## 3.2 软件实现

使用Python语言，sk-learn框架 实现三种算法。使用交叉验证观察准确率，使用框架自带函数构建模型，得到评估参数

主界面如下图3.2.1

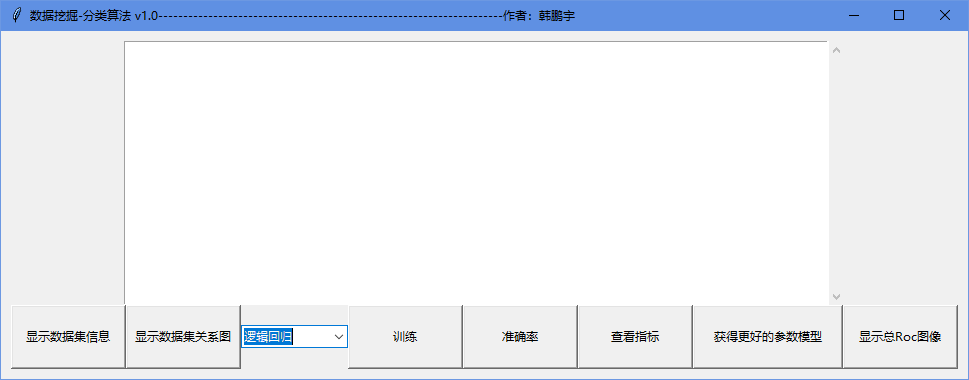


图3.2.1 主界面

### 3.2.1 逻辑回归算法

使用sklearn框架初始化模型：LogisticRegression(solver=**'liblinear'**)，进行训练，得到的指标及准确率如下图3.2.2所示：

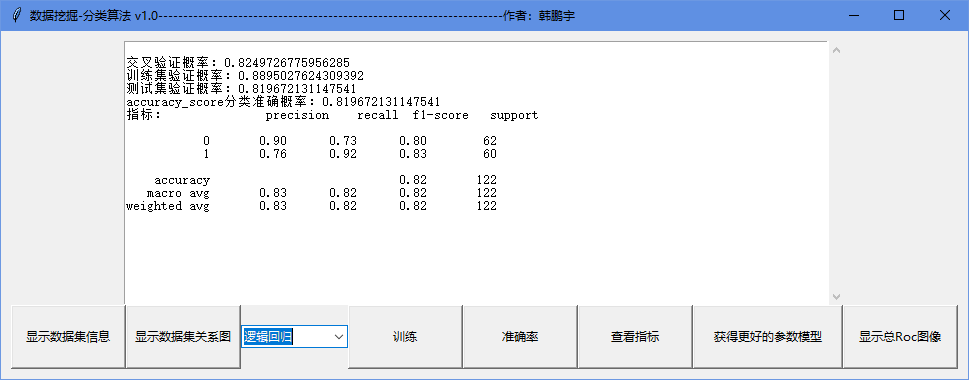


图3.2.2 逻辑回归初始模型

使用网格搜索获得更好的模型：

测试后将 param\_grid = [

{

'C': [0.01, 0.1, 1, 10, 100],

'penalty': ['l2', 'l1'],

'class\_weight': ['balanced', None]

}

]

得到更好的模型后，模型参数准确率如图3.2.3：

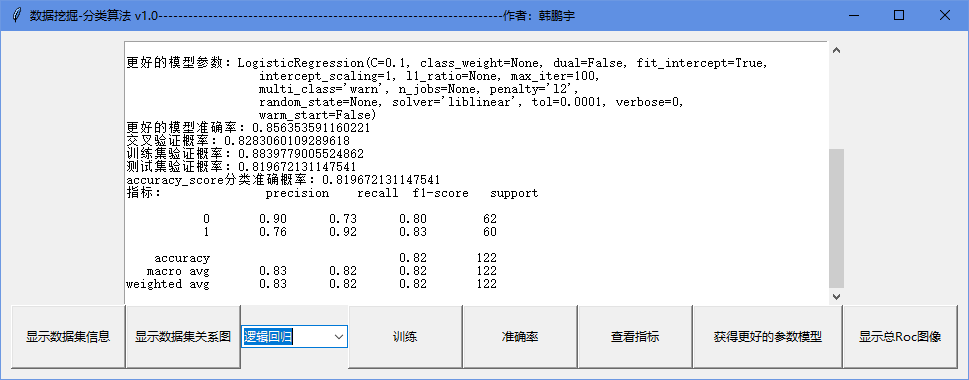


图 3.2.3 更好的逻辑回归模型

再次进行交叉验证得到的概率与上次几乎相同，说明最开始的模型已经最好了。如图3.2.4所示。

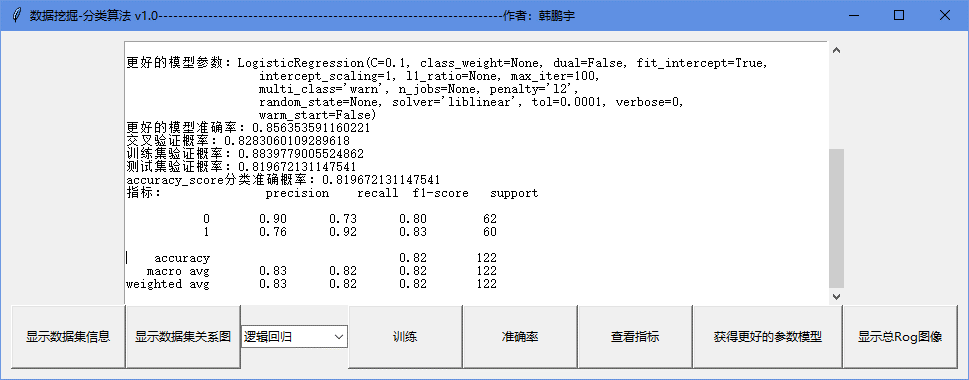


图3.2.4 更好的逻辑回归模型的准确率和指标

### 3.2.2 kNN 算法：

使用sklearn框架的包初始化模型：neighbors.KNeighborsClassifier()，进行训练，得到的交叉验证概率和指标如图3.2.5所示

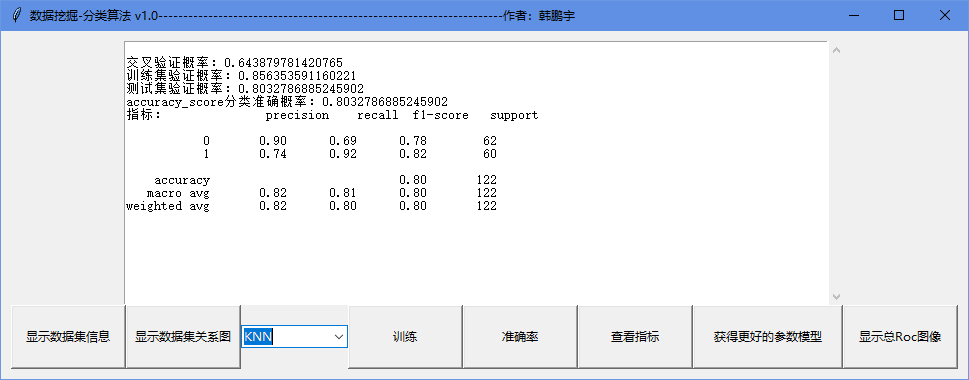


图3.2.5 kNN 初始模型

使用网格搜索获得更好的模型：

测试后将： param\_grid = [

{

'weights': ['uniform'],

'n\_neighbors': [i for i in range(1, 31)]

},

{

'weights': ['distance'],

'n\_neighbors': [i for i in range(1, 31)],

'p': [i for i in range(1, 6)]

}

]

得到更好的模型后，得到的交叉验证概率和指标如图3.2.6图所示：



图3.2.6 kNN更好的模型参数和指标

### 3.2.3 决策树

使用sklearn框架初始化决策树模型：DecisionTreeClassifier()，得到的模型参数和交叉验证概率和指标如图3.2.7所示：

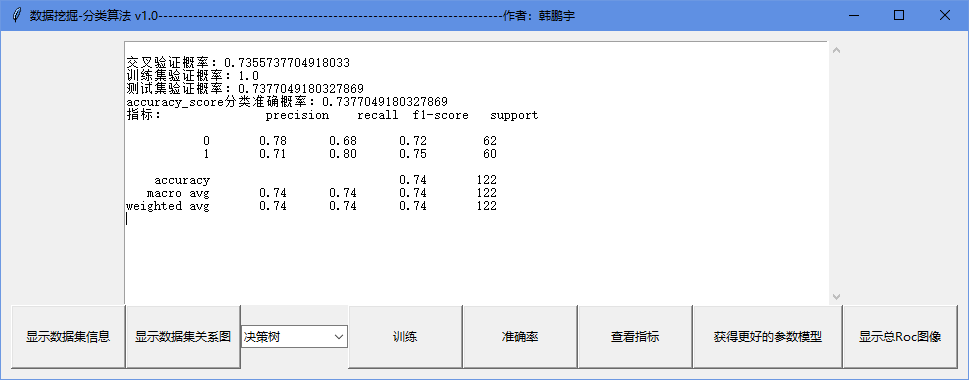


图3.2.7 决策树初始模型

使用网格搜索获得更好的模型参数：

测试后将： param\_grid = [

{

'max\_features': ['auto', 'sqrt', 'log2'],

'min\_samples\_split': [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,16, 17, 18],

'min\_samples\_leaf': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]

} ]

得到更好的模型后，得到的交叉验证概率和指标如图3.2.8图所示：



图 3.2.8 更好的决策树模型

### 3.3结果讨论

使用Roc曲线图像生成，观察三种算法的准确率曲线(初始模型和更好的模型)，如图3.3，3.4所示：

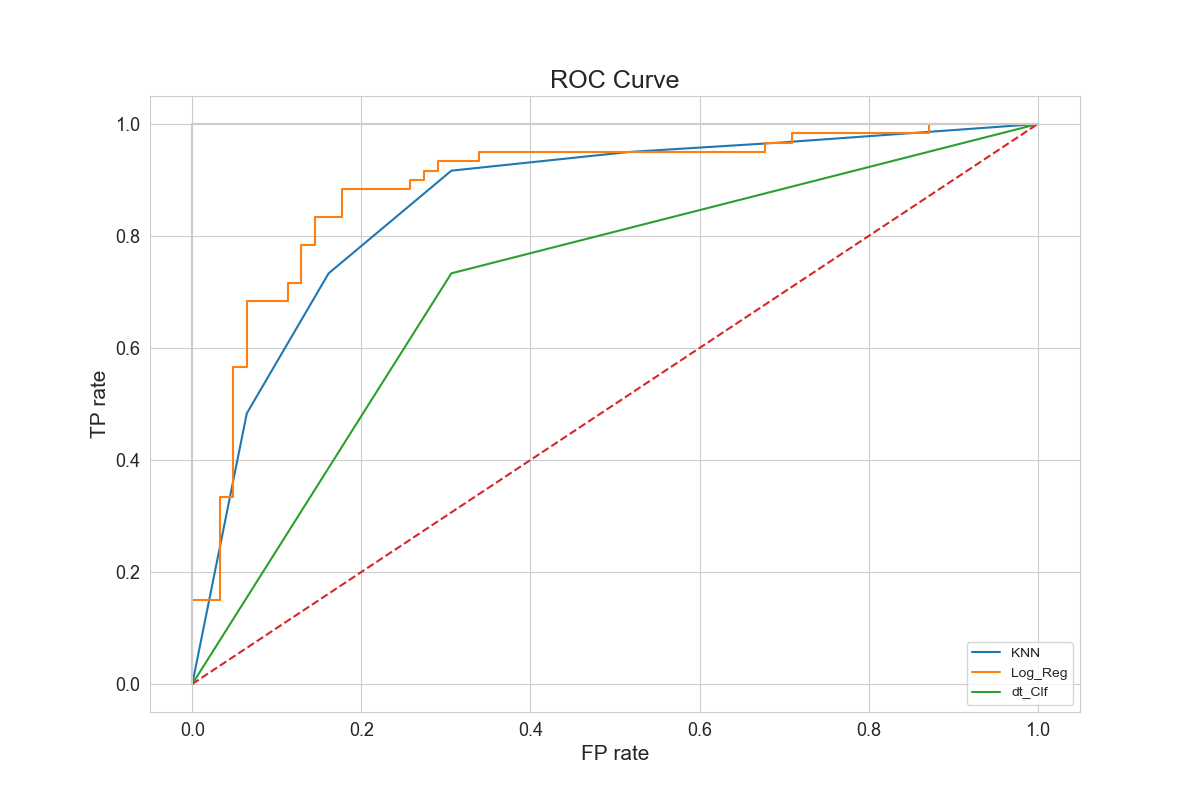


图3.3 初始模型roc图像

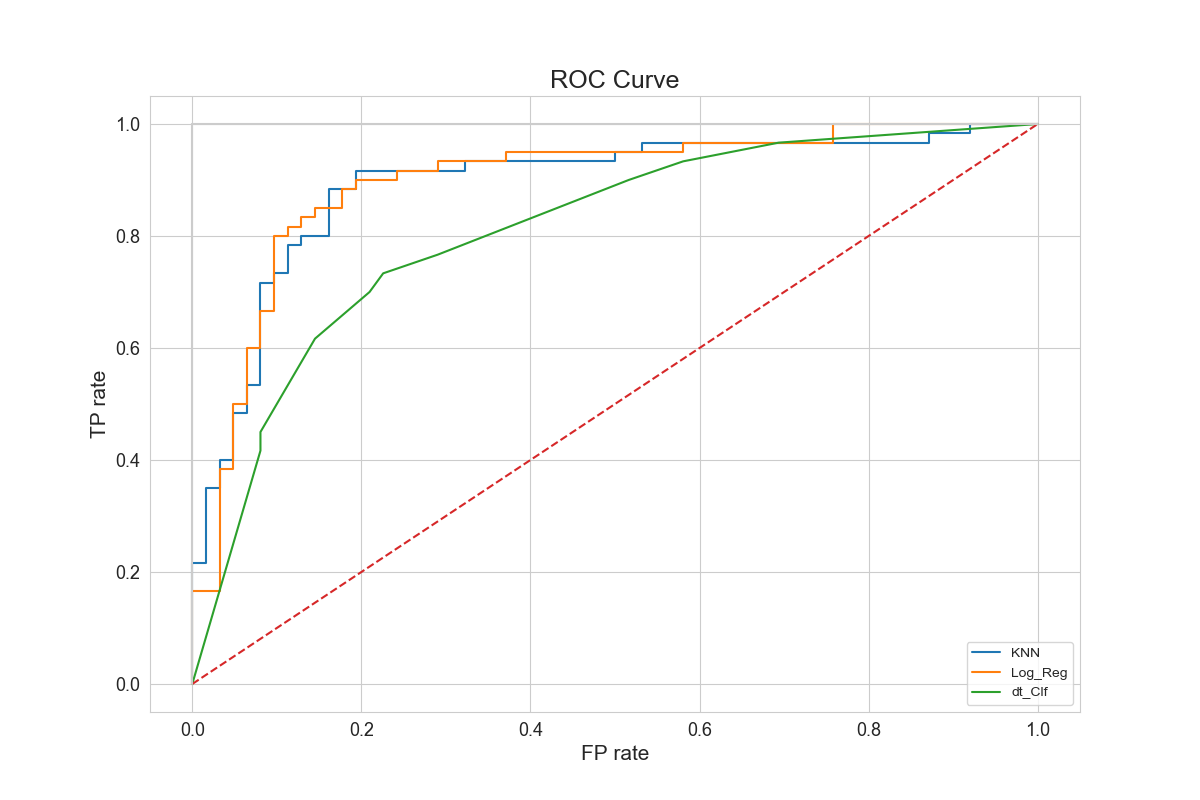


图3.3 三种算法优化后的Roc曲线

由roc曲线和准确率以及模型的指标来看，三种模型中，逻辑回归和kNN对心脏病这个数据集的分类效果更好，而决策树相对来说就一般。

# 4.总结与展望

通过一个学期对数据挖掘课程的学习，我算是接触到了数据挖掘的门槛，也算是刚刚入门，发现了数据的应用大有可为，看似无关的数据，在各种算法上的处理，机器学习后，变得有规律，让我发现了数据挖掘的魅力，希望我在以后的工作学习生活中，能继续钻研数据挖掘算法，继续实践。为祖国的科研事业做出贡献。