

**2022-2023学年第一学期**

**本科生课程作业**

|  |  |
| --- | --- |
| **题目： 目：** | COVID-19 患者死亡的危险因素分析及预测 |

|  |  |
| --- | --- |
| 课程名称 | 多元统计分析 |
| 姓 名 | 杨浩然 |
| 学 号 | 2020116130 |
| 院 系 | 数学学院 |
| 专 业 | 应用统计学 |

**2022年 11月**

目录：

[一. 问题重述 4](#_Toc124521342)

[二. 数据简介 4](#_Toc124521343)

[三.数据集概述 5](#_Toc124521344)

[3.1 导入相关包 5](#_Toc124521345)

[3.2 导入数据 5](#_Toc124521346)

[3.3查看数据基本信息 6](#_Toc124521347)

[3.4检查缺失值并打印每个特征的取值个数 6](#_Toc124521348)

[3.5 查看死亡日期的取值分布 7](#_Toc124521349)

[3.6 查看是否有肺炎的取值分布 7](#_Toc124521350)

[四.数据预处理 8](#_Toc124521351)

[4.1 剔除除“INTUBED”、“怀孕”、“ICU”外的特征缺失值 8](#_Toc124521352)

[4.2预处理“日期死亡”列 8](#_Toc124521353)

[4.3 剔除“怀孕”缺失值 8](#_Toc124521354)

[4.4 INTUBED特征缺失值分析 9](#_Toc124521355)

[4.5 ICU特征缺失值分析 9](#_Toc124521356)

[4.6 打印数据预处理后的每个特征的取值 10](#_Toc124521357)

[五.数据可视化 10](#_Toc124521358)

[5.1 绘制死亡的直方图 10](#_Toc124521359)

[5.2 绘制年龄的直方图 11](#_Toc124521360)

[5.3 绘制年龄-死亡关系图 11](#_Toc124521361)

[5.4 绘制年龄-死亡-性别关系图 12](#_Toc124521362)

[5.5 绘制性别-死亡关系图 12](#_Toc124521363)

[5.6 绘制肥胖-死亡关系图 13](#_Toc124521364)

[5.7 使用热力图可视化特征之间的相关性 13](#_Toc124521365)

[六.解决不平衡数据集问题 14](#_Toc124521366)

[6.1 使用RandomUnderSampler进行欠采样 15](#_Toc124521367)

[6.2 查看欠采样后的数据集 15](#_Toc124521368)

[七.建立逻辑回归模型 15](#_Toc124521369)

[7.1 删除与“死亡”相关性较低的特征 15](#_Toc124521370)

[7.2 将分类特征转换为哑变量 16](#_Toc124521371)

[7.3 使用对离群值具有鲁棒性的统计量缩放功能 16](#_Toc124521372)

[7.4 将数据集分为特征集和标签集 16](#_Toc124521373)

[7.5 划分训练集和测试集 16](#_Toc124521374)

[7.6 建立逻辑回归模型 16](#_Toc124521375)

[7.7 使用F1 Score评估模型 17](#_Toc124521376)

[7.8 使用混淆矩阵评估模型 17](#_Toc124521377)

[7.9 使用ROC曲线评估模型 17](#_Toc124521378)

[7.10 模型优点 18](#_Toc124521379)

[八.建立随机森林模型 18](#_Toc124521380)

[8.1 对欠采样后的数据集进行训练集和测试集的划分 18](#_Toc124521381)

[8.3 建立随机森林模型 18](#_Toc124521382)

[8.4 使用F1 Score评估模型 18](#_Toc124521383)

[8.5 使用混淆矩阵评估模型 19](#_Toc124521384)

[8.6 使用ROC曲线评估模型 19](#_Toc124521385)

[8.7 模型优点 19](#_Toc124521386)

# 一. 问题重述

冠状病毒病（COVID-19）是由一种新发现的冠状病毒引起的传染病。大多数感染 COVID-19 病毒的人会出现轻度至中度呼吸道疾病，无需特殊治疗即可康复。老年人以及患有心血管疾病、糖尿病、慢性呼吸道疾病和癌症等潜在健康问题的人更容易患上严重疾病。

在整个大流行病过程中，医疗保健提供者面临的主要问题之一是医疗资源短缺以及有效分配这些资源的适当计划。在这些困难时期，能够预测一个人在检测呈阳性时或在此之前可能需要什么样的资源，这将对政府有巨大帮助，因为他们将能够采购和安排必要的资源挽救那个病人的生命。

该项目的主要目标是建立一个机器学习模型，根据 Covid-19 患者的当前症状、状态和病史，预测患者是否处于高危状态。

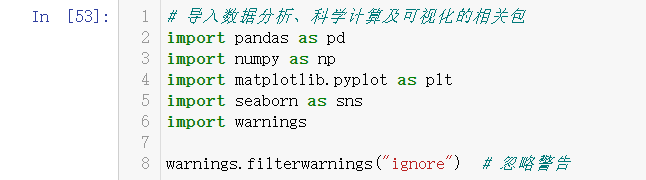
# 二. 数据简介

该数据集由墨西哥政府提供。该数据集包含大量匿名的患者相关信息，包括先决条件。原始数据集包含 21 个独特的特征和 1,048,576 名独特的患者。在布尔特征中，1 表示“是”，2 表示“否”。值 97 和 99 是缺失数据。

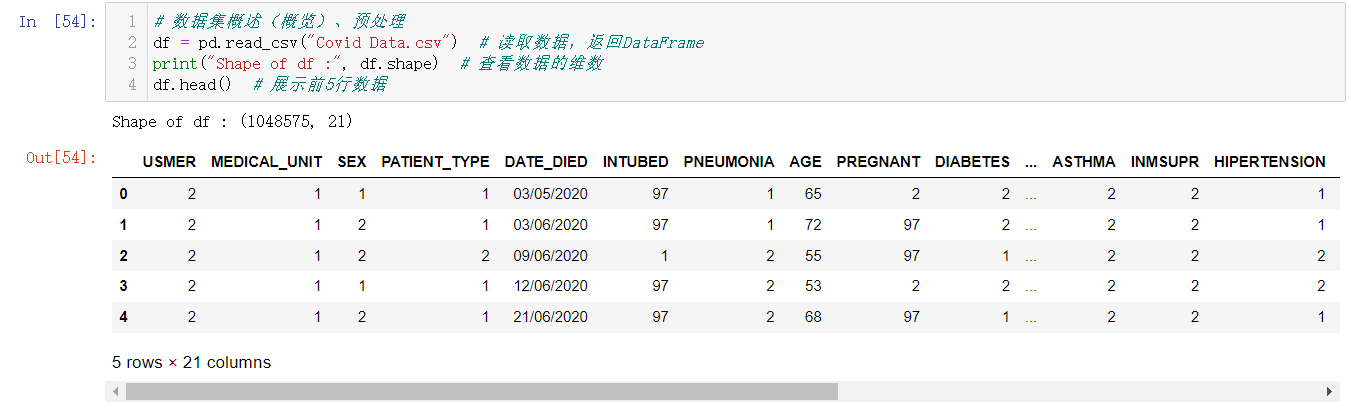
* 性别：1代表女性，2代表男性。
* 年龄：患者。
* 分类：covid 测试结果。值 1-3 表示患者被诊断为不同程度的 covid。4 或更高意味着患者不是 covid 的携带者或测试没有定论。
* 患者类型：患者在单位接受的护理类型。1 人回家，2 人住院。
* 肺炎：患者是否已经有气囊炎症。
* 怀孕：患者是否怀孕。
* 糖尿病：患者是否患有糖尿病。
* copd：表示患者是否患有慢性阻塞性肺疾病。
* 哮喘：患者是否患有哮喘。
* inmsupr：患者是否免疫抑制。
* 高血压：患者是否患有高血压。
* 心血管：患者是否患有心脏或血管相关疾病。
* 肾慢性：患者是否患有慢性肾病。
* 其他疾病：患者是否患有其他疾病。
* 肥胖：患者是否肥胖。
* 烟草：患者是否是烟草使用者。
* usmr：表示患者是否接受过一级、二级或三级医疗单位的治疗。
* 医疗单位：提供医疗服务的国家卫生系统机构类型。
* 插管：患者是否连接到呼吸机。
* icu：表示患者是否已入住重症监护病房。
* 死亡日期：如果患者死亡，请注明死亡日期，否则为 9999-99-99。

# 三.数据集概述

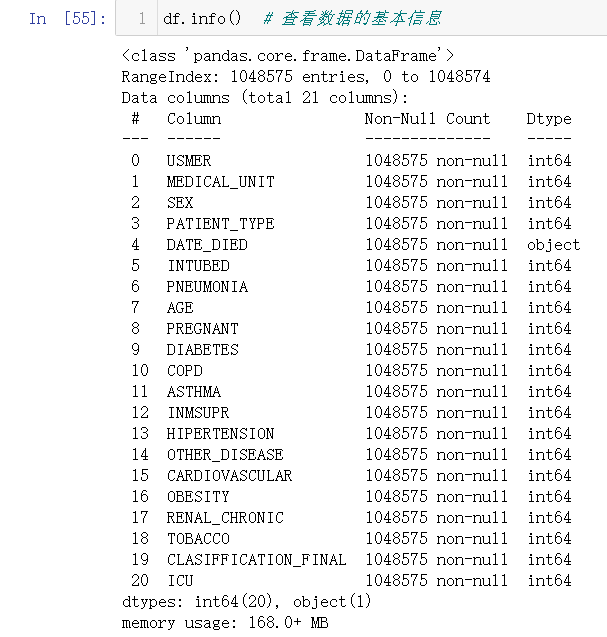
## 3.1 导入相关包



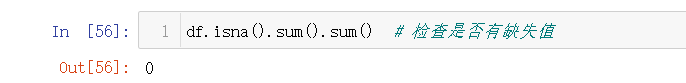
## 3.2 导入数据

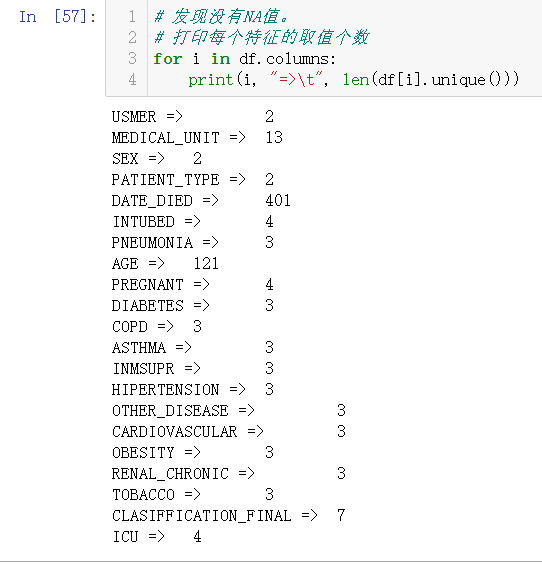


## 3.3查看数据基本信息

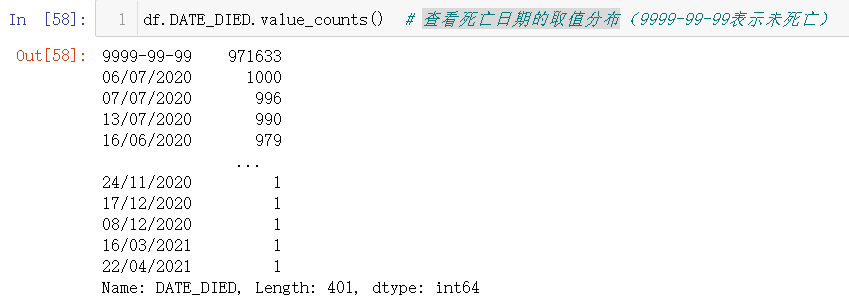


## 3.4检查缺失值并打印每个特征的取值个数

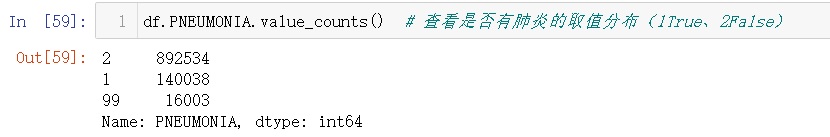




## 3.5 查看死亡日期的取值分布



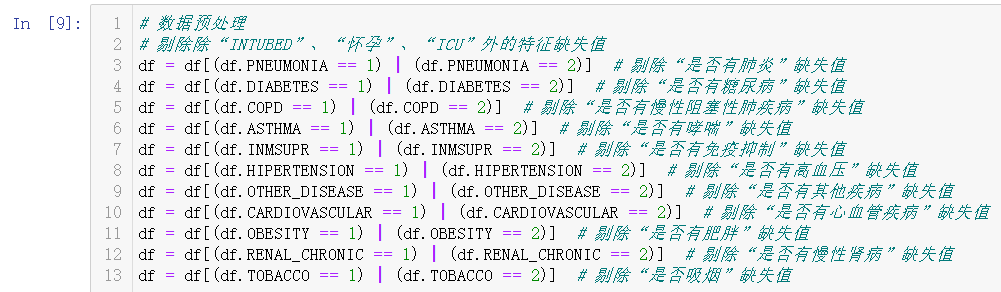
## 3.6 查看是否有肺炎的取值分布



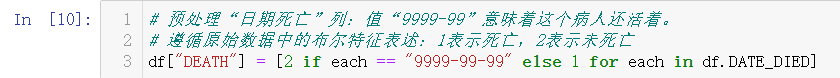
# 四.数据预处理

* 我们有一些功能，我们希望他们只有2个唯一的值。但是我们看到这些特征有3或4个唯一值。例如，"肺炎"功能有3个唯一值（1,2,99）99表示NaN值。因此，我们将只取包含1和2值的行。
* 在“DATE\_DIED”列中，我们有971633个“9999-99”值，代表活着的患者所以我会把这个功能作为一个“死亡”，包括病人是否死亡。

## 4.1 剔除除“INTUBED”、“怀孕”、“ICU”外的特征缺失值

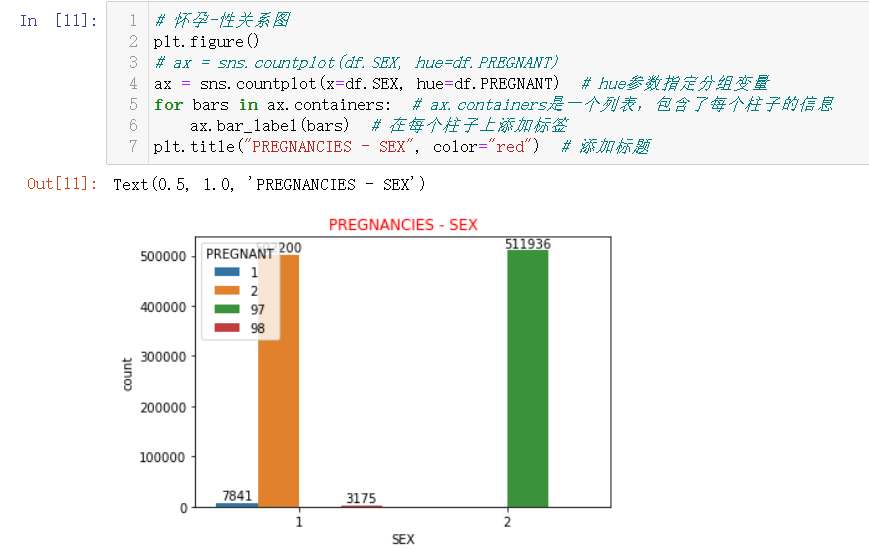


## 4.2预处理“日期死亡”列

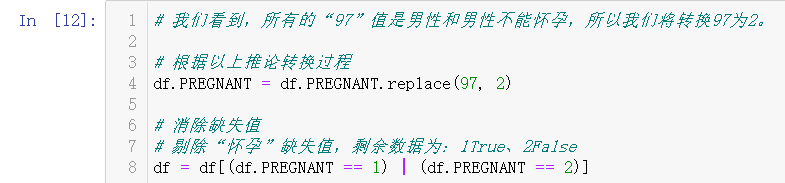


## 4.3 剔除“怀孕”缺失值

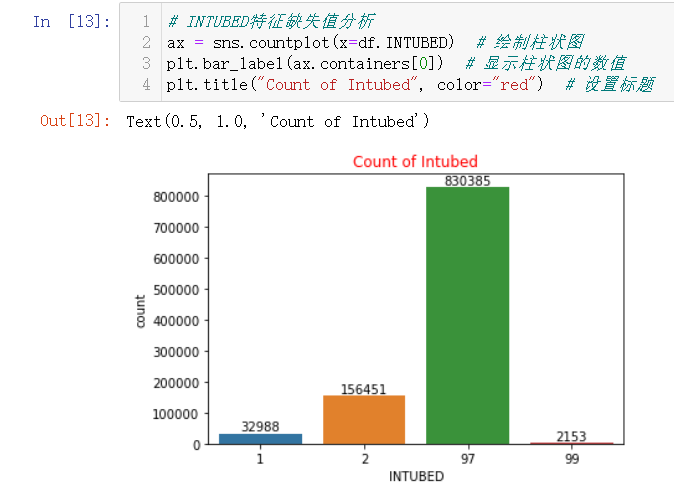
绘制怀孕-性关系图：



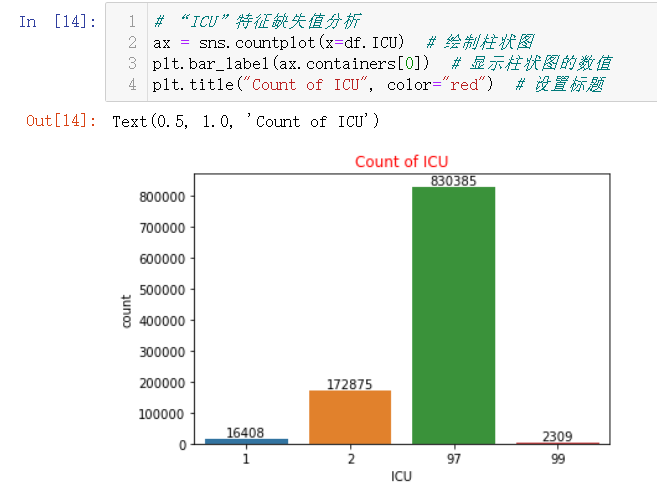
我们看到，所有的“97”值是男性和男性不能怀孕，所以我们将转换97为2。根据以上推论转换过程：



## 4.4 INTUBED特征缺失值分析



## 4.5 ICU特征缺失值分析

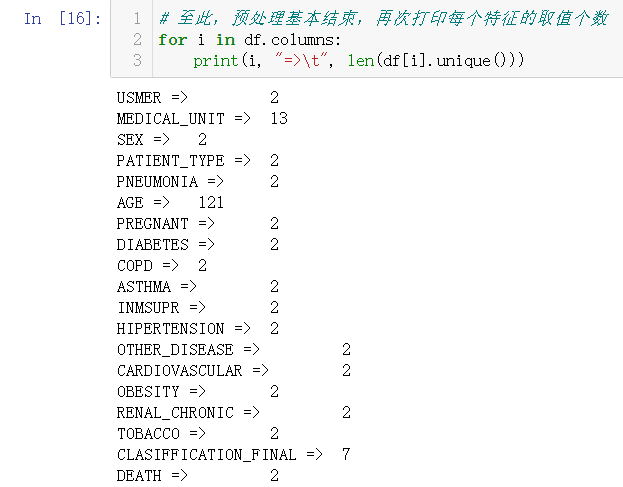


* 在“INTUBED”和“ICU”中有太多缺少的值，下面将删除它们。
* 另外删除的还有“DATE\_DEAD”列，将使用“Death”特征代替。



## 4.6 打印数据预处理后的每个特征的取值

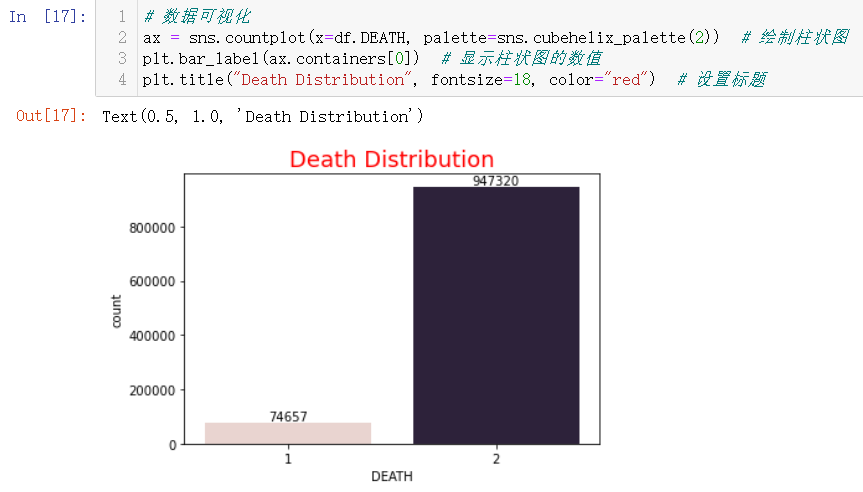
至此，预处理基本结束，再次打印每个特征的取值个数



如上，我们只剩一个被称为“年龄”的数字特征，其余特征都是明确的。

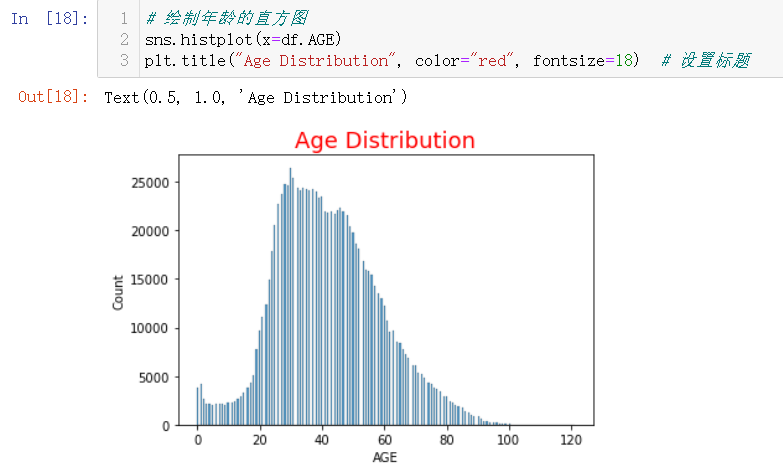
# 五.数据可视化

## 5.1 绘制死亡的直方图



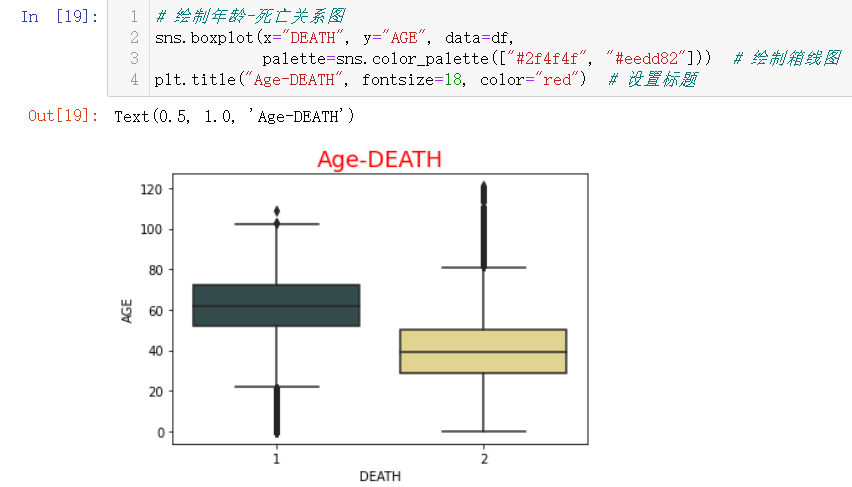
可以看到值在类别（目标）列中的分布并不平衡。这可能会导致模型在测试集合上出现不平衡的问题，这个留到后面再进行解决。

## 5.2 绘制年龄的直方图



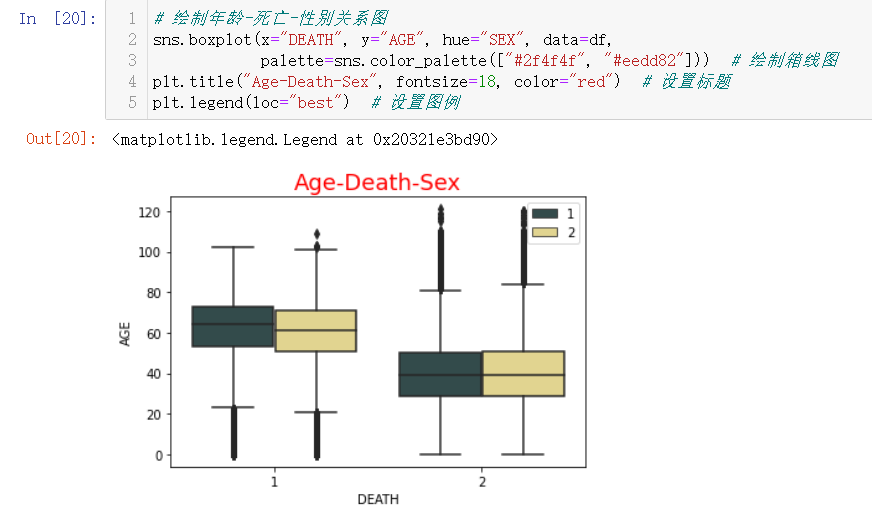
从直方图中可以看出患者大致集中在在20~60岁之间。

## 5.3 绘制年龄-死亡关系图



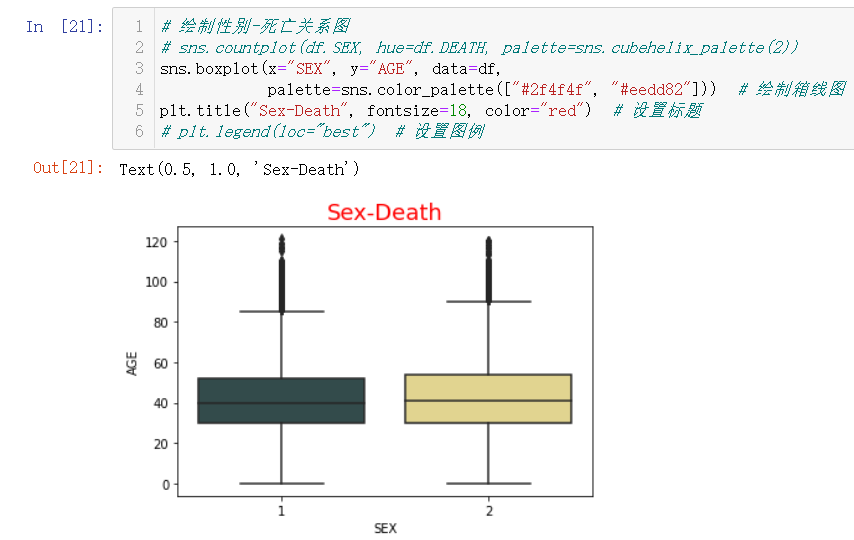
从上面的箱线图我们可以看出老年患者要比年轻患者更容易死亡

## 5.4 绘制年龄-死亡-性别关系图



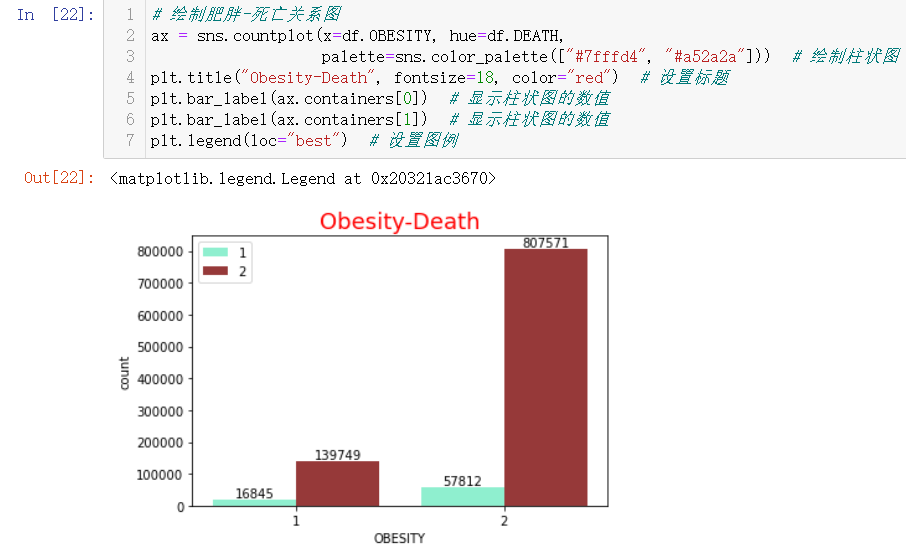
* 老年患者比年轻患者更容易死亡。（上面已经得出过了）
* 从新的箱线图中可以看出病人在平均发病率方面，男女之间没有很大的差别。（下面单独绘制进行校验）

## 5.5 绘制性别-死亡关系图



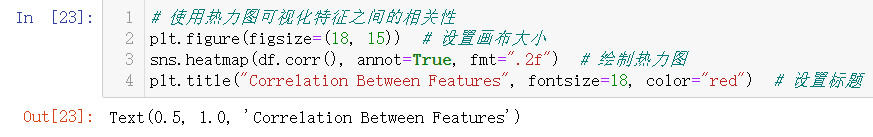
与女性相比，男性更容易死于COVID。（不显著）

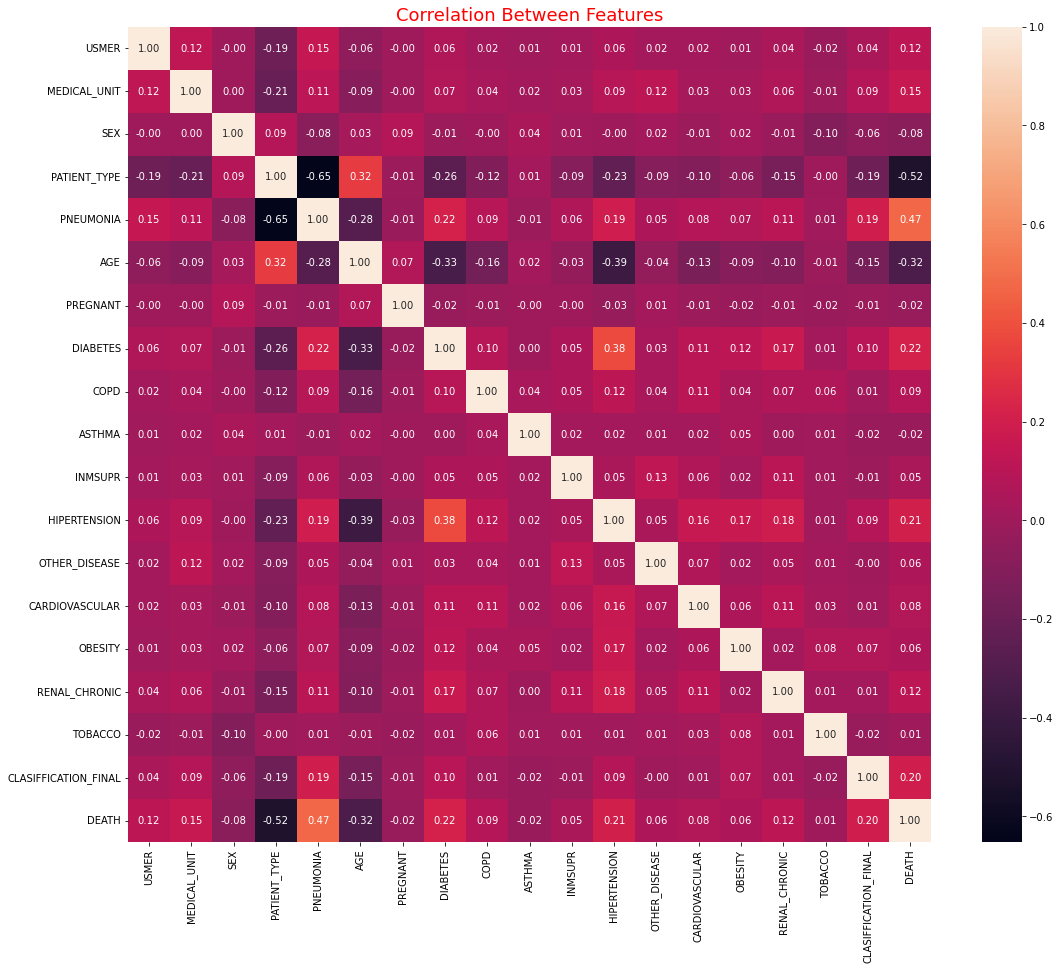
## 5.6 绘制肥胖-死亡关系图



从肥胖-死亡关系柱状图中可以看出肥胖患者比非肥胖患者更容易死于COVID。

## 5.7 使用热力图可视化特征之间的相关性

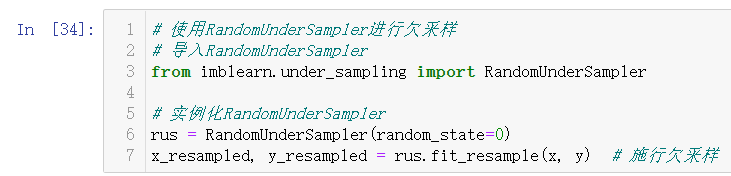




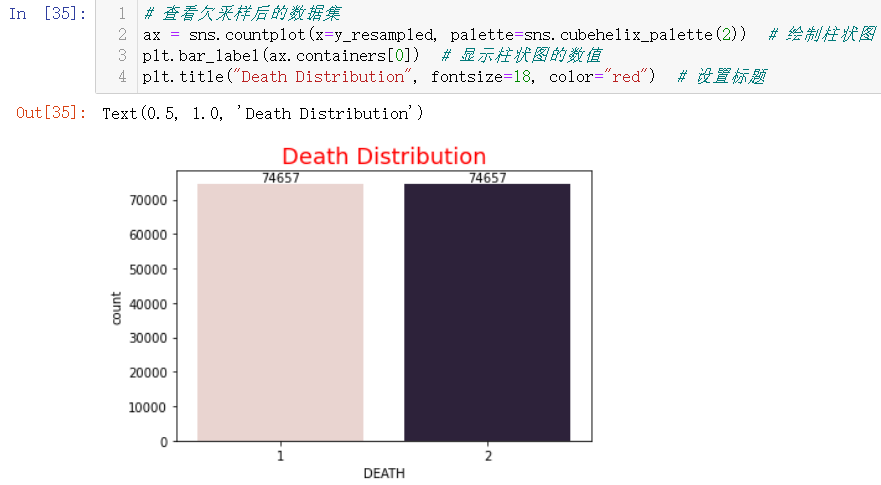
# 六.解决不平衡数据集问题

* 加载更多数据
* 更改性能指标
* 重采样（欠采样或过采样）
  + 欠采样是一种通过保持所有数据在少数类和减少大多数类的大小来平衡不均衡的数据集的技术。
  + 下面将使用欠采样的方式尝试解决数据集的不平衡问题，之所以选择欠采样是因为我们拥有很多的病人数据，如果使用过采样将增加太多数据。
  + 如果使用欠采样仍然没有解决问题再考虑其他。
* 改变算法
* 惩罚模型等。

## 6.1 使用RandomUnderSampler进行欠采样

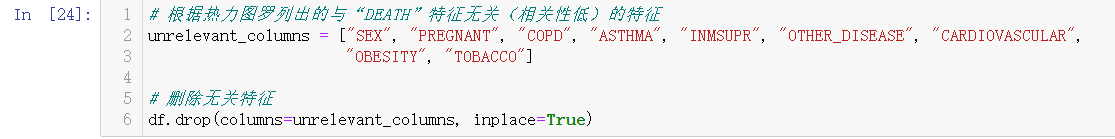


## 6.2 查看欠采样后的数据集

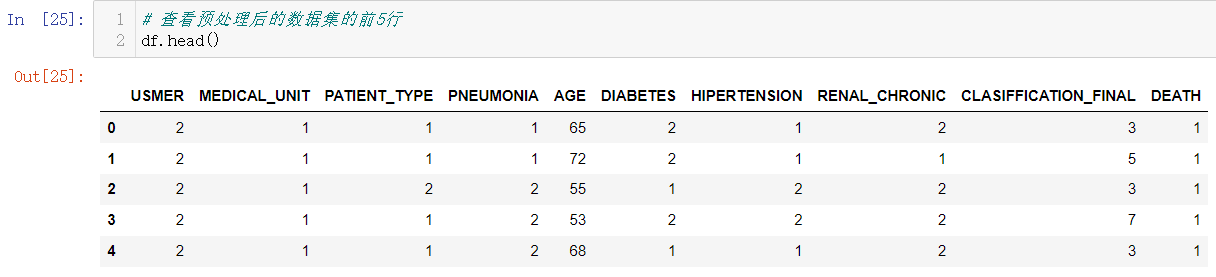


# 七.建立逻辑回归模型

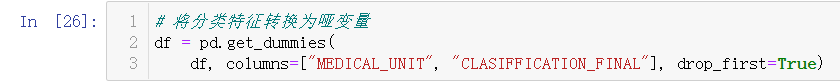
## 7.1 删除与“死亡”相关性较低的特征



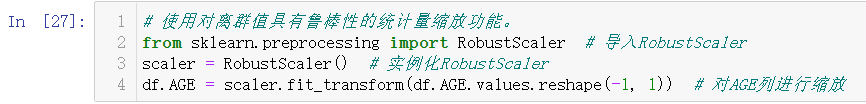
查看预处理后的数据集的前5行



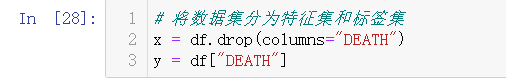
## 7.2 将分类特征转换为哑变量



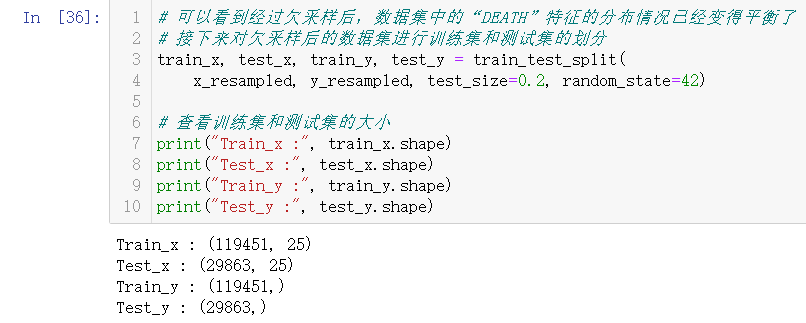
## 7.3 使用对离群值具有鲁棒性的统计量缩放功能



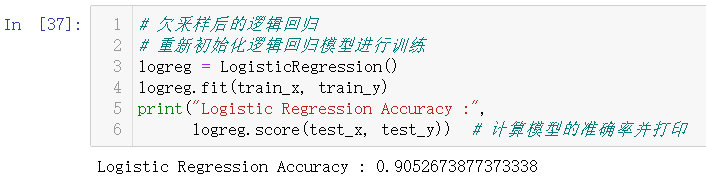
## 7.4 将数据集分为特征集和标签集



## 7.5 划分训练集和测试集

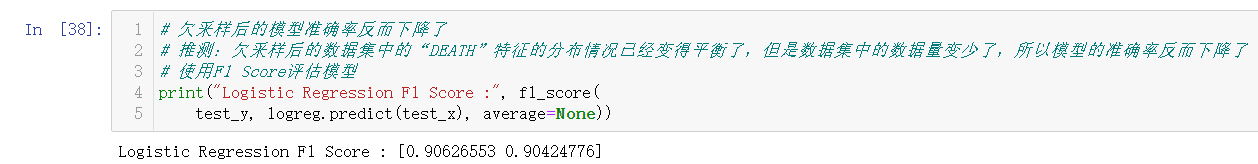


## 7.6 建立逻辑回归模型

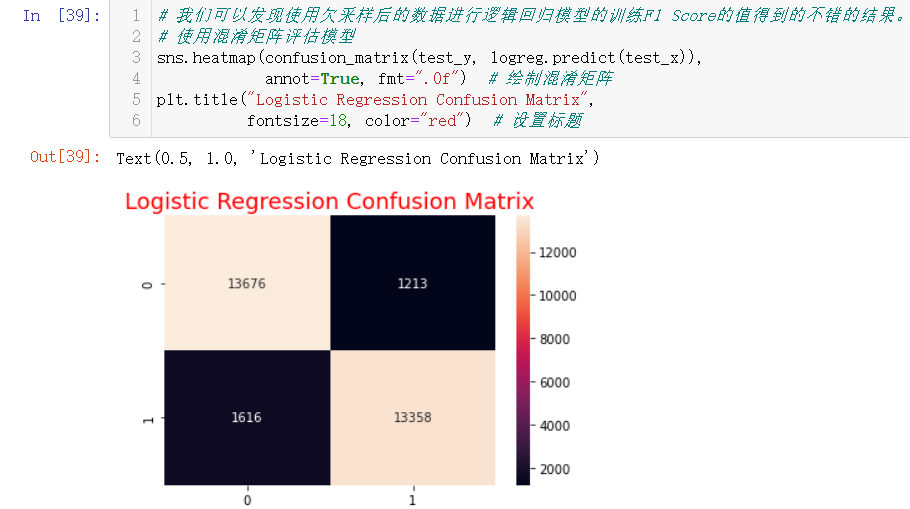


可以看出使用逻辑回归分析的准确性较好。但它可能会误导我们，所以接下来将检查其他指标。

## 7.7 使用F1 Score评估模型

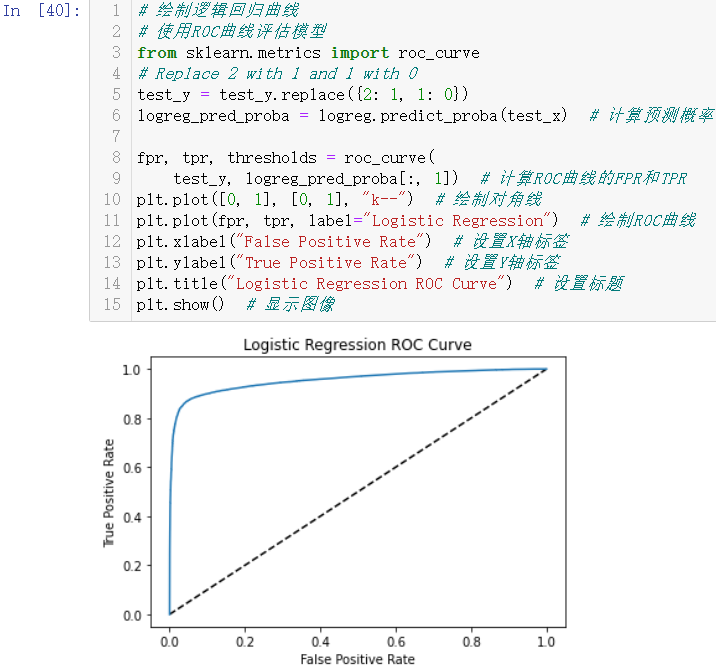


## 7.8 使用混淆矩阵评估模型



通过F1和混淆矩阵发现使用欠采样可以很好解决前面提到过数据集分类特征的不均衡可能导致问题出现。

## 7.9 使用ROC曲线评估模型

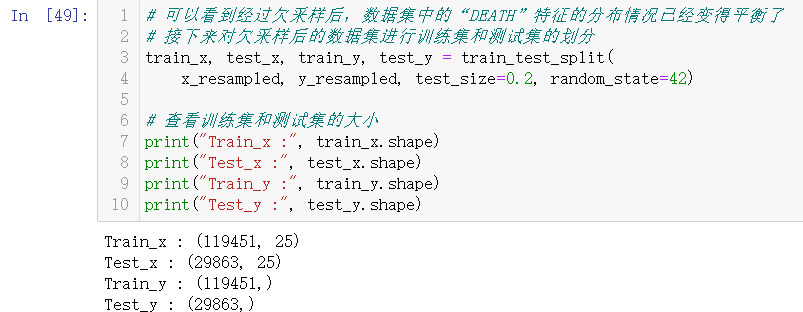


## 7.10 模型优点

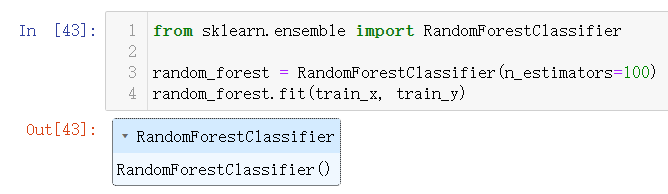
* 本模型使用逻辑回归对于“死亡”的预测情况，准确率为90.52%。
* 使用欠采样算法解决了不平衡的数据集问题。

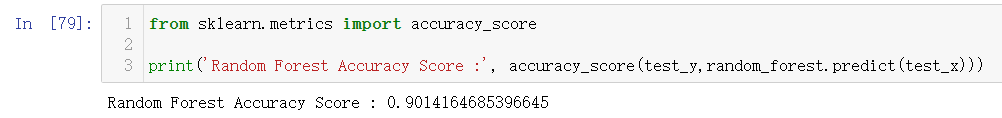
# 八.建立随机森林模型

## 8.1 对欠采样后的数据集进行训练集和测试集的划分

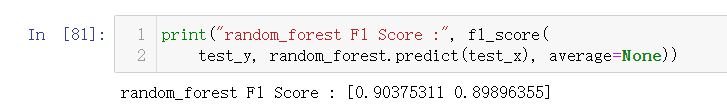


## 8.3 建立随机森林模型

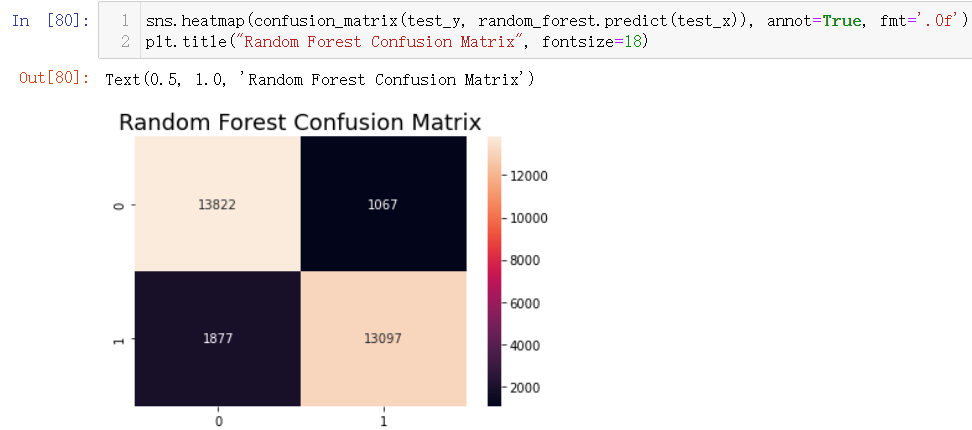




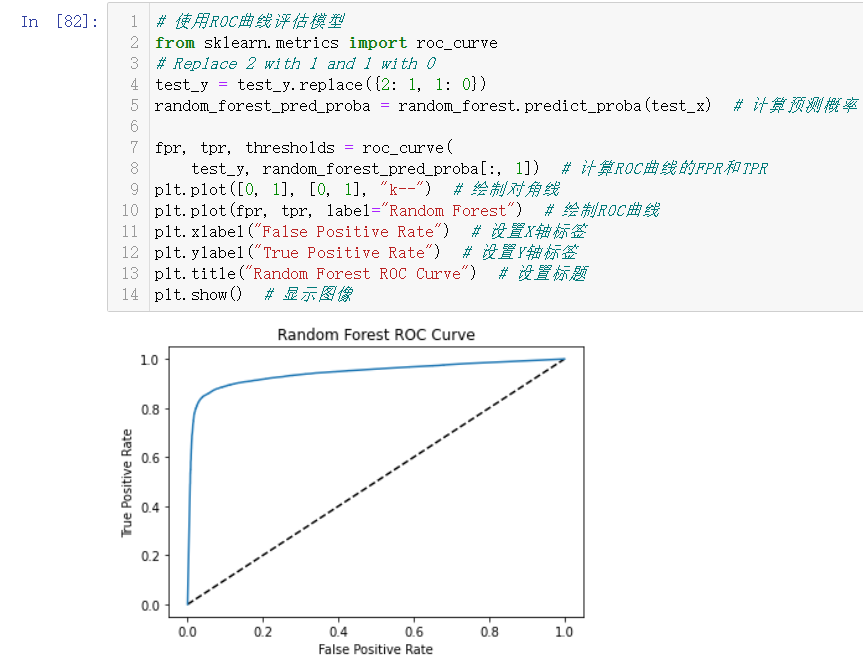
## 8.4 使用F1 Score评估模型



## 8.5 使用混淆矩阵评估模型



## 8.6 使用ROC曲线评估模型



## 8.7 模型优点

* 本模型使用逻辑回归对于“死亡”的预测情况，准确率为90.14%。
* 使用欠采样算法解决了不平衡的数据集问题。