UCI 心脏病数据集数据分析报告

曹竞轩、张子琪、徐克丰、殷超、冯小芝

一、背景描述

医疗大数据是指个人从出生到死亡的全生命周期过程中，由免疫、体检、门诊、住院等健康活动所产生的大数据。医疗大数据的数据来源有四类：临床医学数据、药理研究和生命科学数据和个人健康数据。通过对医疗大数据的分析、加工，可以挖掘出在疾病诊断、治疗、公共卫生防治等方面的重要价值。

其中，健康医疗大数据是一种高附加值的信息资产，虽然个体健康医疗数据对于医疗技术革新的价值有限，但通过对海量、来源分散、格式多样的数据进行采集、存储、深度学习和开发，可以从中发现新知识、创造新价值、提升新能力，从而进一步反哺健康医疗服务产业。因此，健康医疗大数据的发展关乎国计民生，具有重大的战略性意义。

现如今，医疗水平已经得到飞速提升，但是有关心脏的疾病依旧是人类面临的医学难题之一。根据最新发布的《中国心血管病报告 2018》，中国心血管病患病率及死亡率仍处于上升阶段。据推算，我国心血管病现患人数为 2.9 亿，死亡率居首位，占居民疾病死亡构成的 40%以上；美国心脏病协会 2018 年发布的《心脏病与卒中统计数据》同样显示，心脏相关的疾病是美国国民的第一大死因，每 7 人中就有 1 人死于冠心病。

因此，我们认为心脏病的诊断和预测是现如今医疗产业中不容忽视的问题。利用数据科学研究方法能有助于我们深入了解心脏病，并为解决心脏病这一医学难题所有助益。

二、数据概述

我们选择的数据是 UCI 机器学习库中的克利夫兰心脏病数据集。本数据集共有 303 个观察样本。与心脏病有关的因素很多，该数据集选用了与患心脏病相关程度较高的 14 个变量，前 13 个可以看作是导致心脏病发作的自变量，第 14 个（诊断结果）可以看作是因变量。这些变量中，既有定性变量，又有定量变量。

在原数据集中，定性变量多用 1，2，3……等数字来表示不同类别，使得数据显得更加简洁直观。通过对不同变量之间关系的研究，可以得出一些与心脏病相关的变量及其相关程度。

各个变量的含义如下表所示：

变量说明表		
变量名		说明与相关评述
age	年龄	反映个体的年龄。 年龄是导致心脏病或心血管疾病的一个重要因素，根据临床经验，每过十年患病的风险会增加大约 3 倍。
sex	性别	反映个体的性别。 男性比绝经前的女性更易患心脏病，而过了更年期，女性的患病风险可能与男性相同。并且如果女性患有糖尿病，她可能比男性更容易患心脏病疾病。因此性别是一个重要变量。
cp	胸痛类型	反映个体的胸痛类型，1=典型心绞痛，2=非典型心绞痛，3=非心绞痛，4=渐进式疼痛。 心绞痛是在心肌没有得到足够的富氧血液的情况下由胸痛或不适引起的。
trestbps	静息血压	反映个体静息血压值（以 mmHg 为单位）。 长时间高血压会损害供养心脏的动脉。高血压和其它症状并发会导致个人健康有很大风险。
chol	血清总胆固醇	反映血清总胆固醇（以 mg/dl 为单位）。 高水平的低密度脂蛋白胆固醇最有可能导致动脉狭窄，而高密度脂蛋白胆固醇会降低心脏病发作的风险。
fbs	空腹血糖	将个体空腹血糖浓度与 120mg/dl 比较，>120mg/dl 记为 1，<=120mg/dl 记为 0。 分泌的胰岛素不足或者胰腺对胰岛素没有反应会导致身体的血糖水平会升高，增加患心脏病的风险。
restecg	静息心电图	反映静息心电图结果，0=正常，1=ST-T 波异常，2=左心室肥大。 对于患心血管疾病较低的人来说，静息心电图得出的结论是可靠的，但

		目前证据表明，对于中度到高风险的人群来说不能充分评估筛查的利弊。
thalach	最大心率	反映个体达到的最大心率。 心血管的风险增加与心率加速有着密切联系，会导致患高血压的风险增加。
exang	运动诱发心绞痛	1=是，2=否。心绞痛导致的疼痛或不适一般包括感觉紧绷、紧握或挤压，可能轻微可能严重。
oldpeak	ST 段下降	运动引起的相对于休息状态的的 ST 段下降，用整数或浮点数来表示。运动的 ECGs 与上倾斜的 ST 段凹陷通常被认为是“模棱两可”的测试。一般情况下，在较低的工作量（以 METs 计算）或较低的心率情况下，出现水平或向下倾斜的 ST 段下凹，表明预后差和多血管病变的可能性高。另一个暗示显著的 CAD 是 ST 段抬高>1mm（通常表明跨壁缺血），这样的病人经常急需冠状动脉造影术。
slope	运动高峰 ST 段	1=抬高，2=平缓，3=下降。 （评述和补充说明同上）
ca	主要血管数	在透视中观察到的的主要血管数量，取值为 0，1，2，3
thal	地中海贫血	3=正常，6=固定缺陷，3=正常缺陷。 地中海贫血是一组遗传性溶血性贫血疾病。由于遗传的基因缺陷致使血红蛋白中一种或一种以上珠蛋白链合成缺如或不足所导致的贫血或病理状态。各种脏器由于铁质的沉积会有一定的损害，导致心力衰竭。
target	诊断结果	反映个体是否患有心脏疾病，0=正常，1，2，3，4=患病

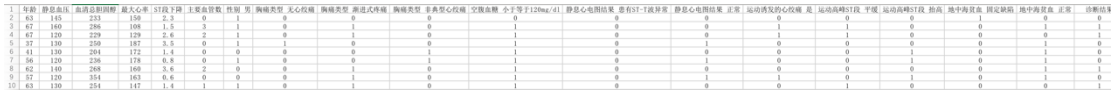
三、数据预处理

在原数据集的 303 个样本中，有 6 个样本有缺失值，数目不大，所以我们将其直接删除。

对余下的样本，我们首先将所有列的列名从缩写改为其完整含义。接下来，有一些定性特征的类别在原数据集中用数字表示，我们将其改为类别名。更改完成后，将新数据集导出为 heart_disease_preprocessed.csv 文件，如下图所示。所有余下的分析将基于这个预处理后的新数据集完成。

	年龄	性别	胸痛类型	静息血压	血清总胆固醇	空腹血糖	静息心电图结果	最大心率	运动诱发的心绞痛	ST段下降	运动高峰ST段	主要血管数	地中海贫血	诊断结果
2	63	男	典型心绞痛	145	233	大于120mg/dl	左心室肥大	150	否	2.3	下降	0	固定缺陷	健康
3	67	男	渐进式疼痛	160	286	小于等于120mg/dl	左心室肥大	108	是	1.5	平缓	3	正常	患病
4	67	男	渐进式疼痛	120	229	小于等于120mg/dl	左心室肥大	129	是	2.6	平缓	2	可逆缺陷	患病
5	37	男	无心绞痛	130	250	小于等于120mg/dl	正常	187	否	3.5	下降	0	正常	健康
6	41	女	非典型心绞痛	130	204	小于等于120mg/dl	左心室肥大	172	否	1.4	抬高	0	正常	健康
7	56	男	非典型心绞痛	120	236	小于等于120mg/dl	正常	178	否	0.8	抬高	0	正常	健康
8	62	女	渐进式疼痛	140	268	小于等于120mg/dl	左心室肥大	160	否	3.6	下降	2	正常	患病
9	57	女	渐进式疼痛	120	354	小于等于120mg/dl	正常	163	是	0.6	抬高	0	正常	健康
10	63	男	渐进式疼痛	130	254	小于等于120mg/dl	左心室肥大	147	否	1.4	平缓	1	可逆缺陷	患病

除此之外，在建模分析时，所有定性数据需要被转换为定量数据。我们采用了添加哑变量的做法，将有 n 个类别的特征拆分为 n-1 个独立的 0-1 二分类特征。n-1 个特征中第 i 个取值为 1，说明原特征的类别为第 i 类；全部取值为 0，说明原特征的类别为第 n 类。所有建模工作将基于经进一步处理的新数据集完成，以下不再赘述。



四、描述分析

图 0——数据概览

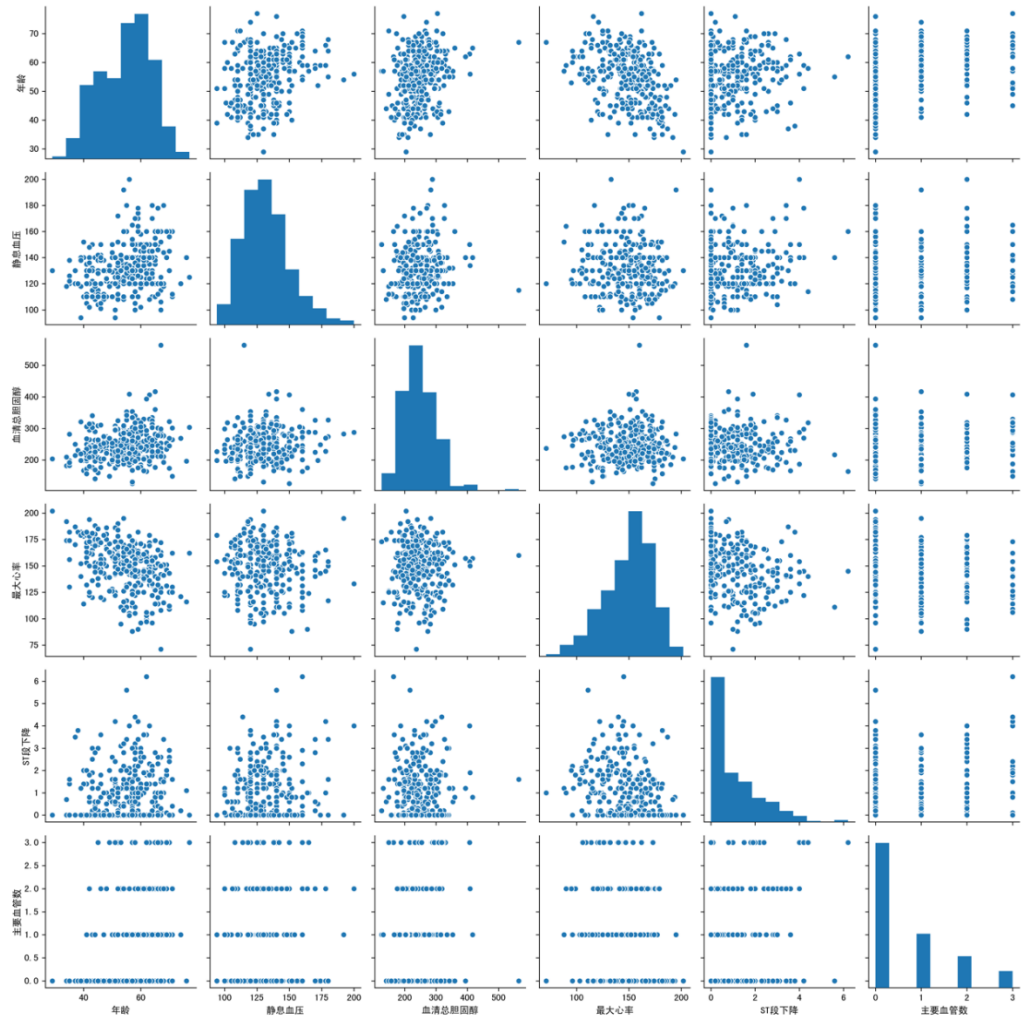
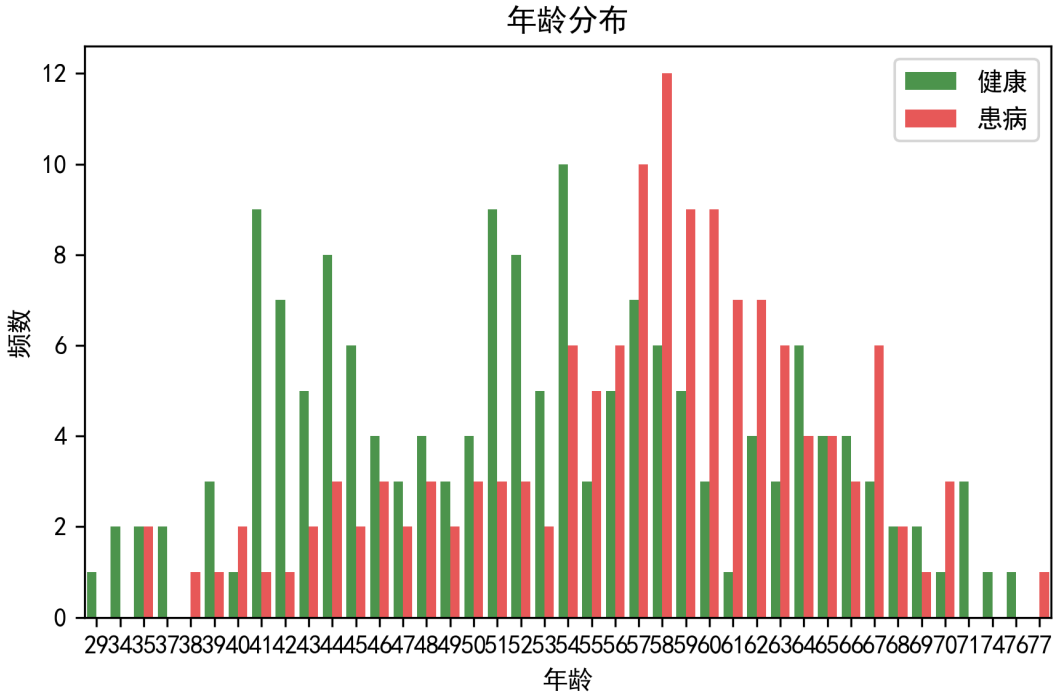
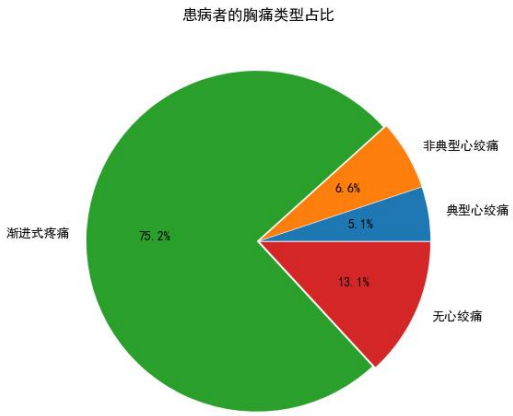


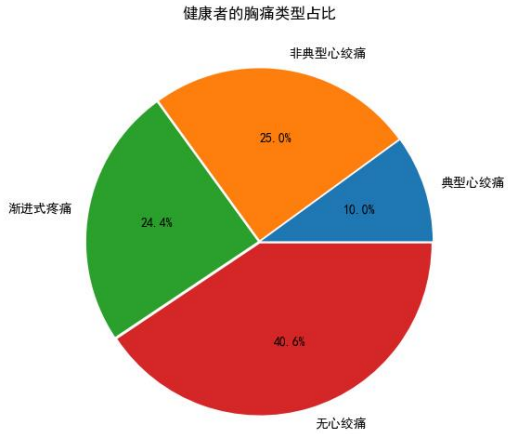
图 1——患病与健康的年龄分布条形图



以年龄为自变量，健康/患病人数为因变量，绿色代表健康，红色代表患病，绘制出条形图。以横轴 54 岁为分界线，可以看出在 54 岁及以前，健康者的数量普遍大于患者数量，在 54 岁以后，患者的数量普遍大于健康者数量。从而可以推断，年龄大的人群患心脏病的概率更大。

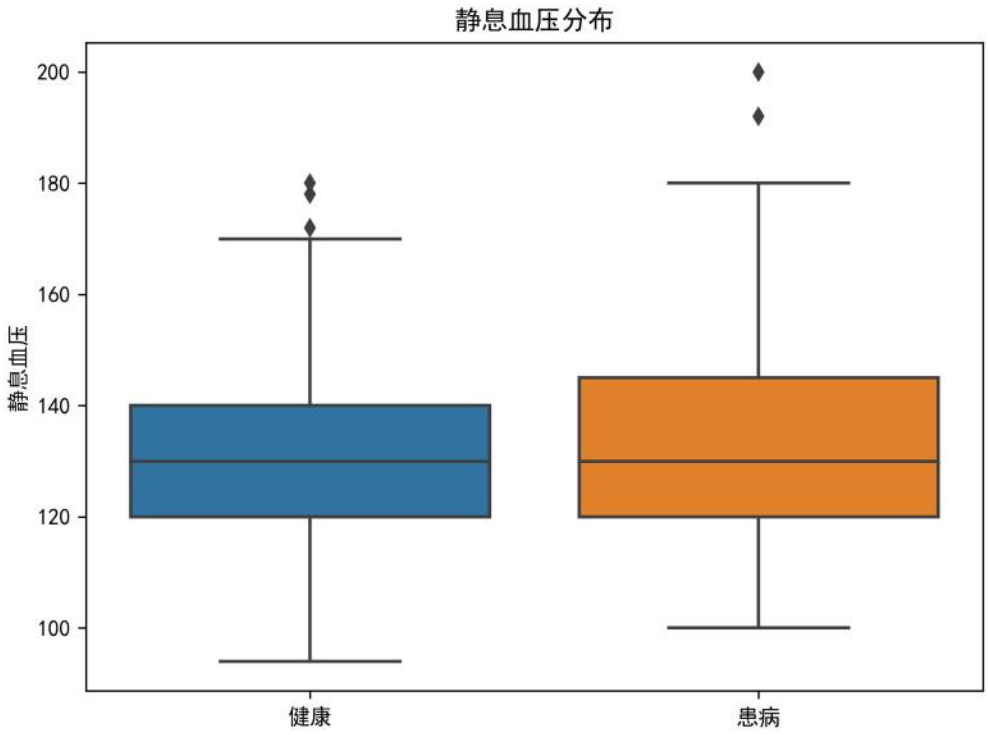
图 2，图 3——患病/健康的胸痛类型饼状图





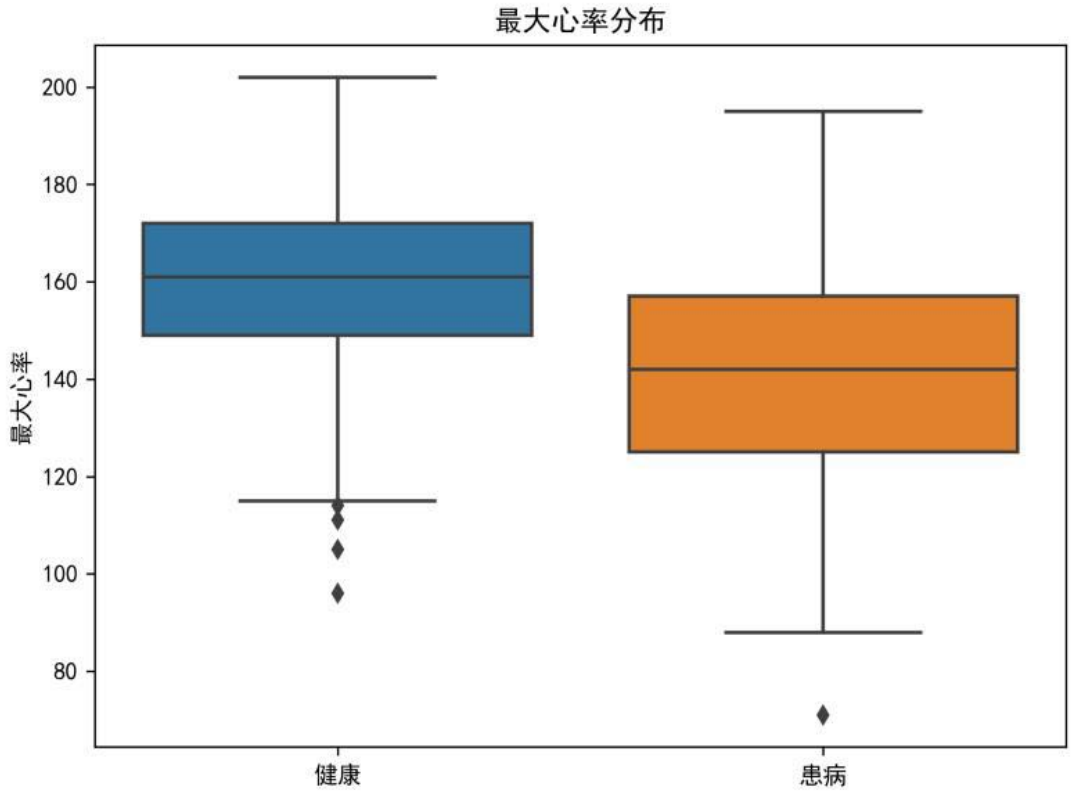
以胸痛类型为自变量，以人数占比为因变量，不同颜色代表不同的胸痛类型，分别对健康者和患者数据绘制出饼图。从图二可以看出，患者的胸痛类型大多为渐进式疼痛，其他疼痛类型或无疼痛均较少；从图三可以看出，健康者中尽管无心绞痛者占比较大，但占比也不到 50%，渐进式疼痛和非典型心绞痛都达到了 20%以上。可以推断，心脏病患者的胸痛类型以渐进式疼痛为主，健康者无心绞痛者居多。

图 4——患病/健康的静息血压箱线图



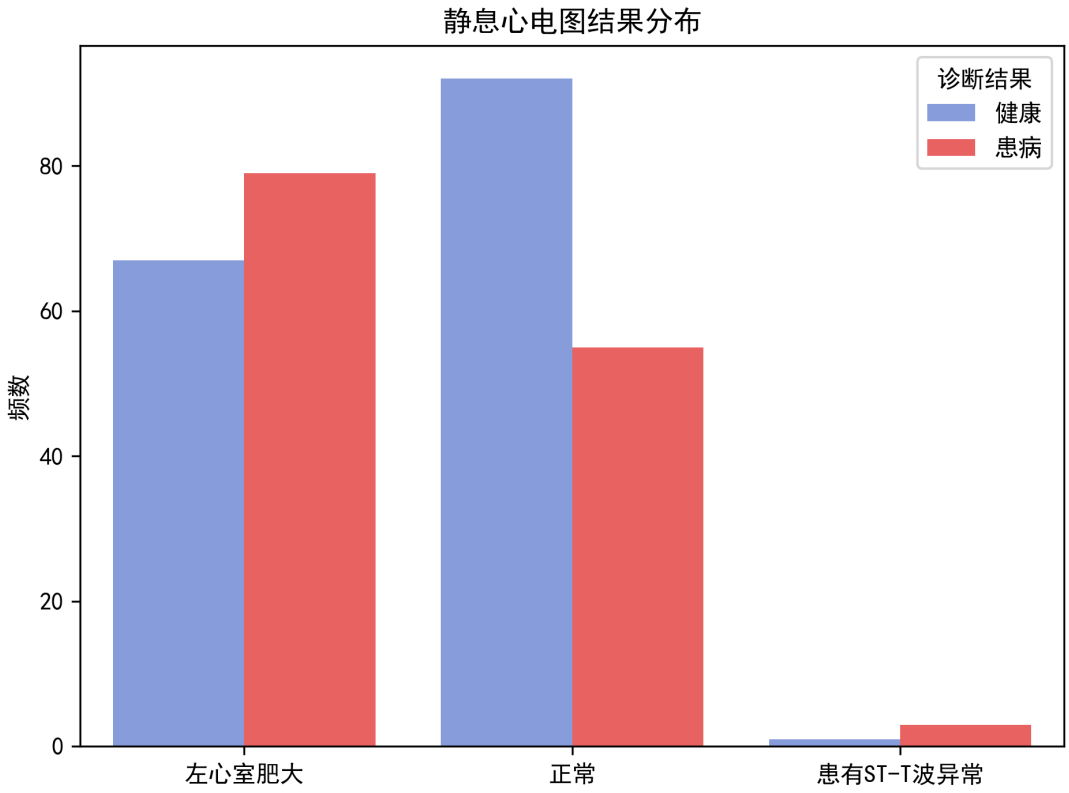
以静息血压为特征，蓝色和橙色分别代表健康者和患者。图中患者静息血压分布的箱线图的四分位数都比健康者的相应数据高，并且患者中存在静息血压极高的 outlier。可以推断，一般来说，心脏病患者的静息血压高于健康者的静息血压。

图 5——患病/健康的最大心率分布箱线图



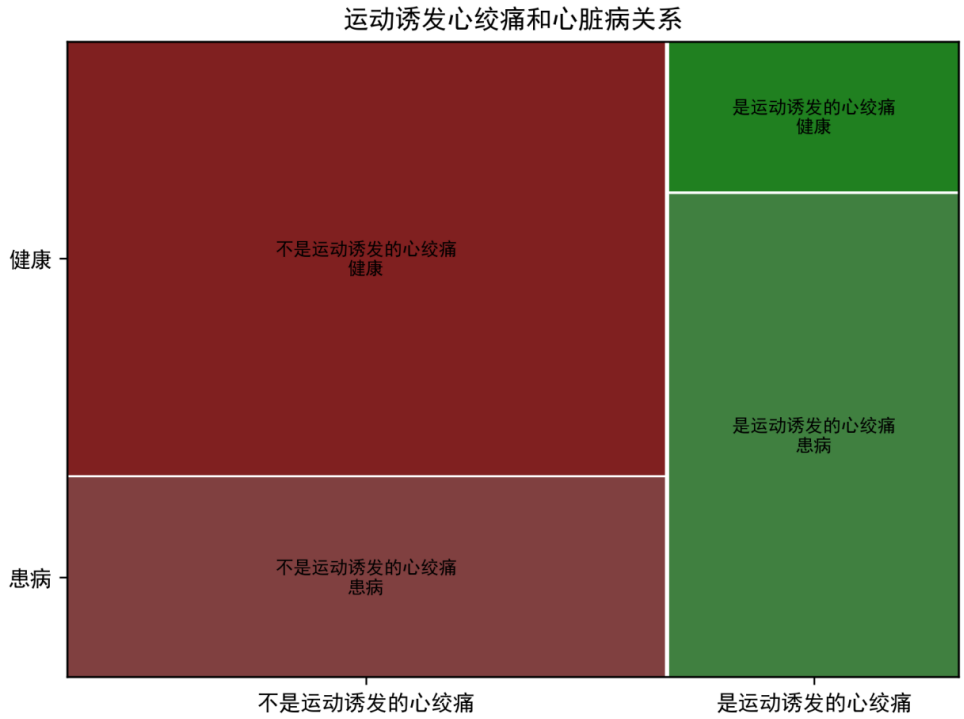
以最大心率为特征，蓝色和橙色分别代表健康者和患者。从图中可见健康者最大心率分布的四分位数都显著高于患者相应数据，并且患者数据中存在极低的异常值，我们可以推断，普遍情况下，患者最大心率低于健康人。

图 6——患病/健康的静息心电图的条形图



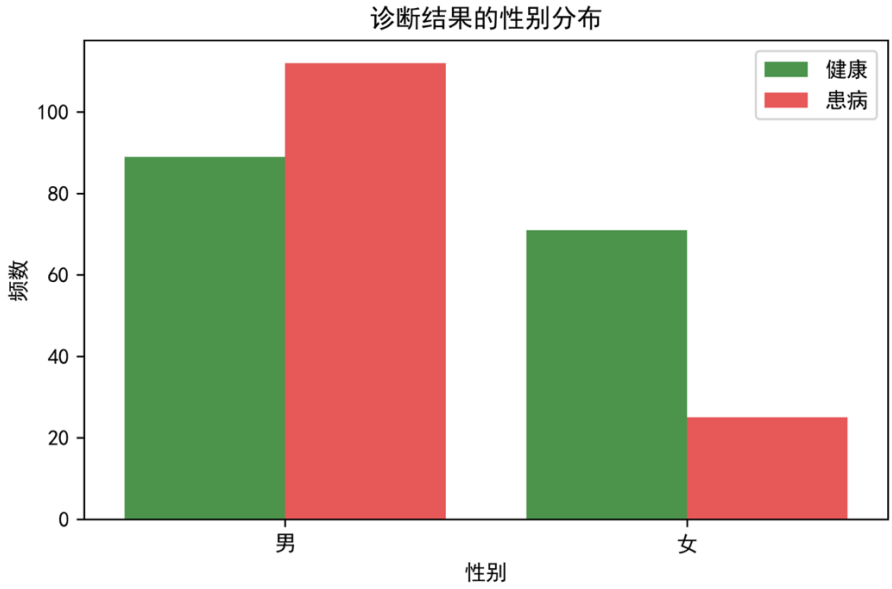
图中紫色、红色分别表示诊断结果分别为健康和患病的人数。通过逐一比较我们可以看出，健康人群中静息心电图结果为左心室肥大人群低于患病人群，健康人群中静息心电图结果显示正常结果高于患病人群，健康人群中静息心电图结果为患有 ST-T 波异常同样低于患病人群。由此我们可以推断，通过静息心电图结果可以在一定程度上区分健康人群和患病人群，但可能存在缺陷。并且静息心电图结果为是否患有 ST-T 波异常以及是否正常能较好的区分健康人群和患病人群。

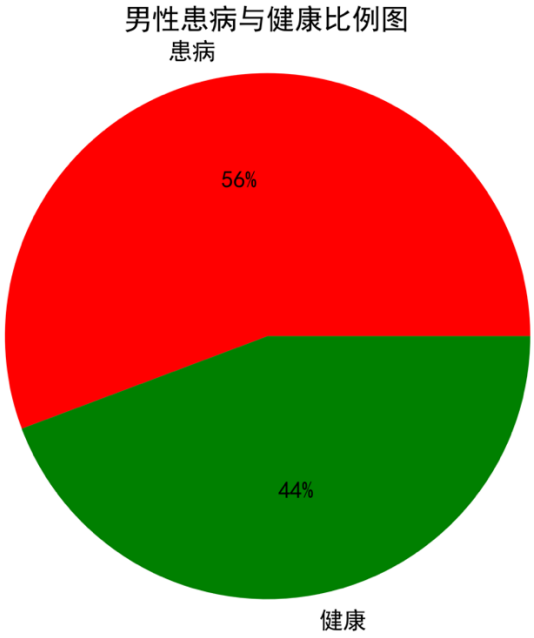
图 7——运动诱发心绞痛类型与心脏病的马赛克图

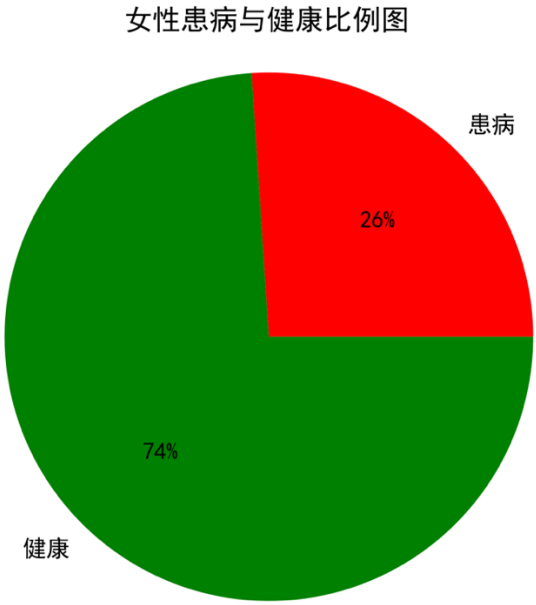


图中每一块面积代表健康或者患病人群中是否为运动诱发心绞痛的频数，其中红色和绿色分别代表不是运动诱发心绞痛和是运动诱发的心绞痛，通过对比我们可以看出，患者中是运动诱发的心绞痛占比显著高于健康人群中运动诱发的心绞痛占比，由此推断，患者更容易因为运动而诱发心绞痛。

图 8，图 9，图 10——男性和女性的患病与健康条形图和饼状图

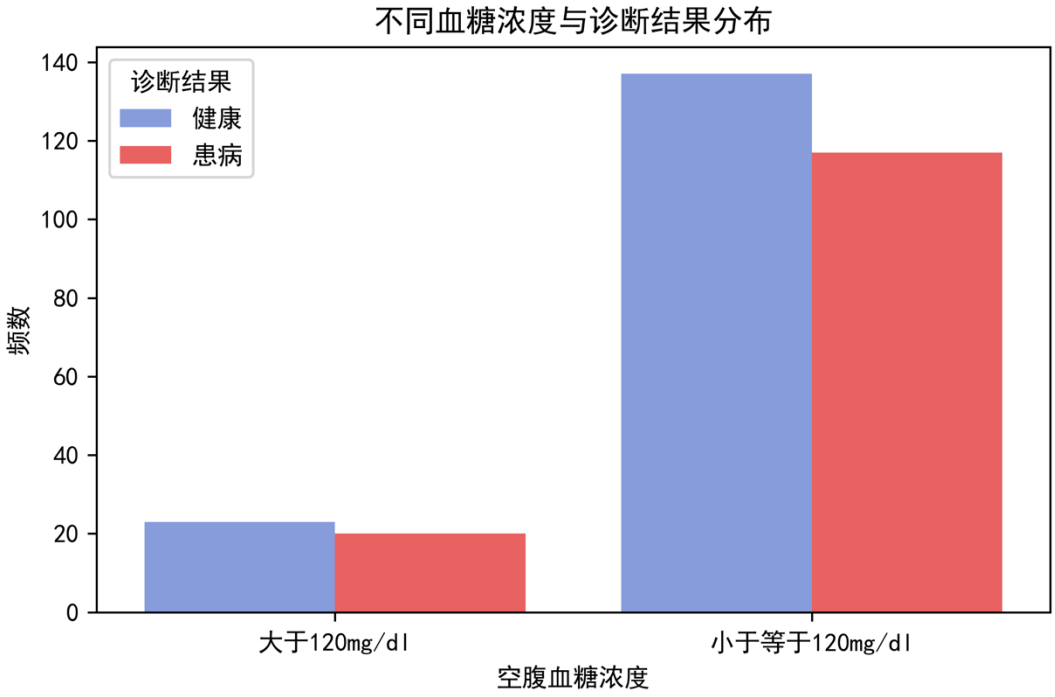


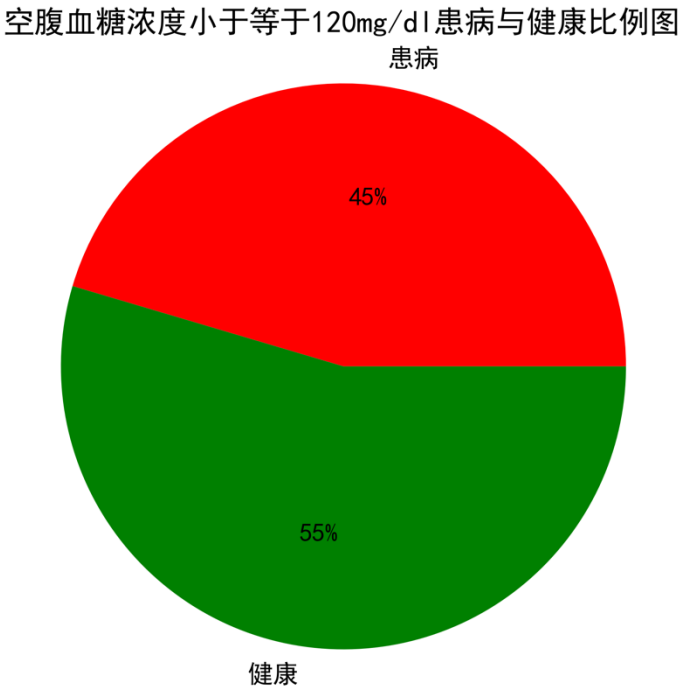




三张统计图中绿色均代表健康，红色均代表患病，这与日常的认知是一致的。其中，在图 8 条形图中，以性别为自变量，分别统计了男性和女性的健康和患病人数。而在图 9 和图 10 中，分别统计了男性，女性患病与健康的所占比例。无论是从条形图还是饼状图中，我们都可以比较直观地得出结论：在该样本中，男性患病比例明显高于女性。

图 11，图 12，图 13——不同血糖浓度与是否患病的条形图和饼状图





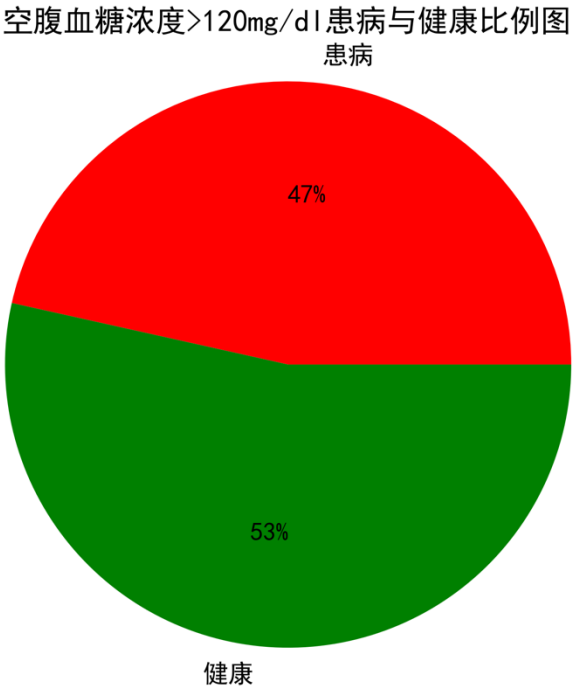


图 11 条形图中，自变量是空腹血糖浓度，不同颜色代表诊断结果。而图 12 和图 13，分别是空腹血糖浓度小于等于，大于 120mg/dl 的诊断结果比例图。由于空腹血糖浓度大于 120mg/dl 的人数很少，这里条形图的显示效果不如比例图。但可以得出的结论是，在这个样本中，空腹血糖浓度和是否患病没有太大联系。

图 14——不同诊断结果的血清总胆固醇箱线图

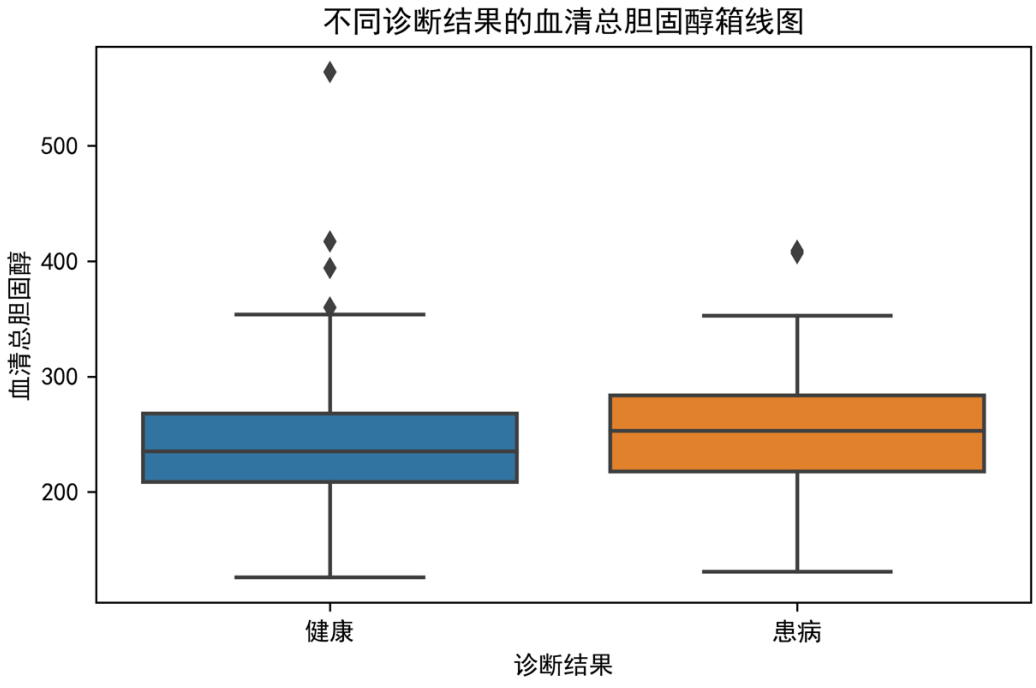


图 14 是箱线图，自变量是诊断结果，而因变量是血清总胆固醇含量。观察两个箱线图可以发现，样本中，健康的人的血清总胆固醇含量的平均值和两个四分位点都比患病者偏低，但是图中健康类的箱线图 Outliers 比较多。

图 15——胆固醇平均值对比

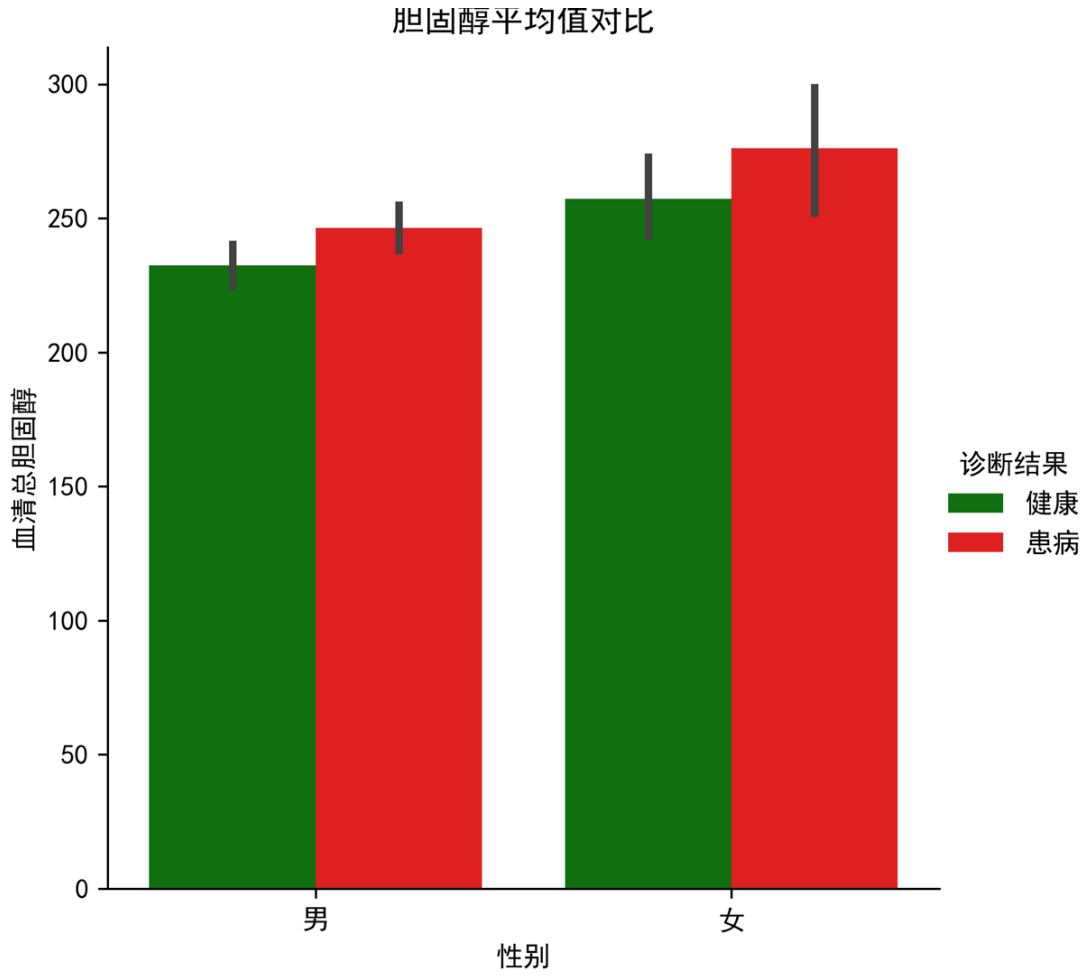


图 15 反映了多维数据。横轴为性别变量，不同颜色代表不同诊断结果，而纵轴为血清总胆固醇含量，柱状图的高度反映了某类人血清胆固醇含量的平均值(这里样本是确定的，所以柱状图顶部的黑色竖线不用考虑)。从图中我们可以发现，无论是健康的人还是患病的人，女性的胆固醇平均含量都高于男性；而无论是男性还是女性，患病者的平均胆固醇含量都要高于健康的人。

图 16——年龄与血清胆固醇散点图

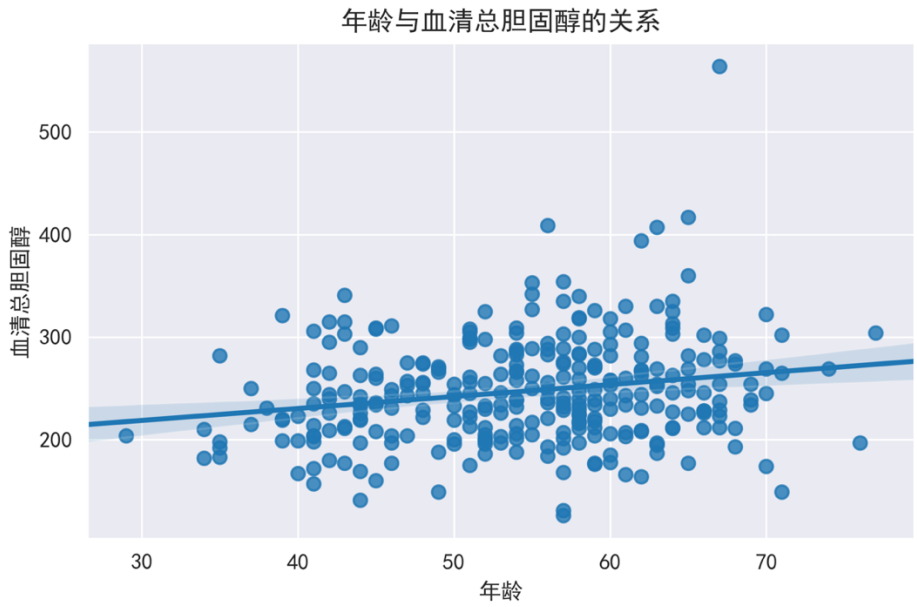


图 16 以年龄为自变量，血清总胆固醇为因变量，画出了散点图，通过直线拟合，发现年龄与血清总胆固醇之间存在一定程度的线性关系，年龄越高，一般来说，血清总胆固醇的含量越高。

图 17——年龄与血清总胆固醇及确诊散点图

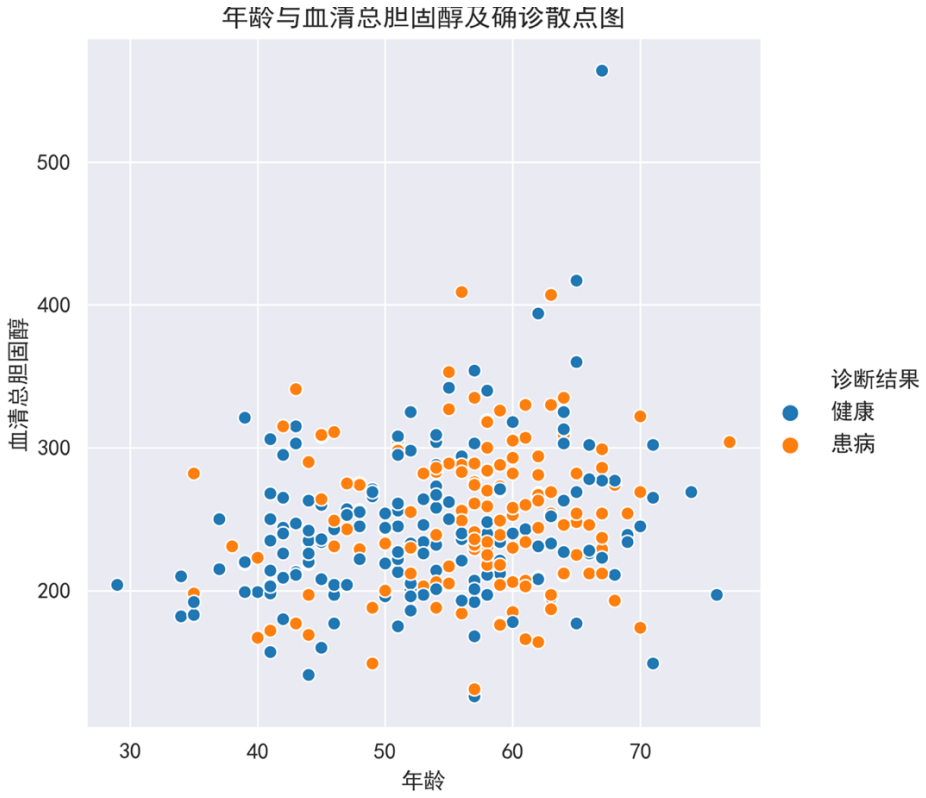
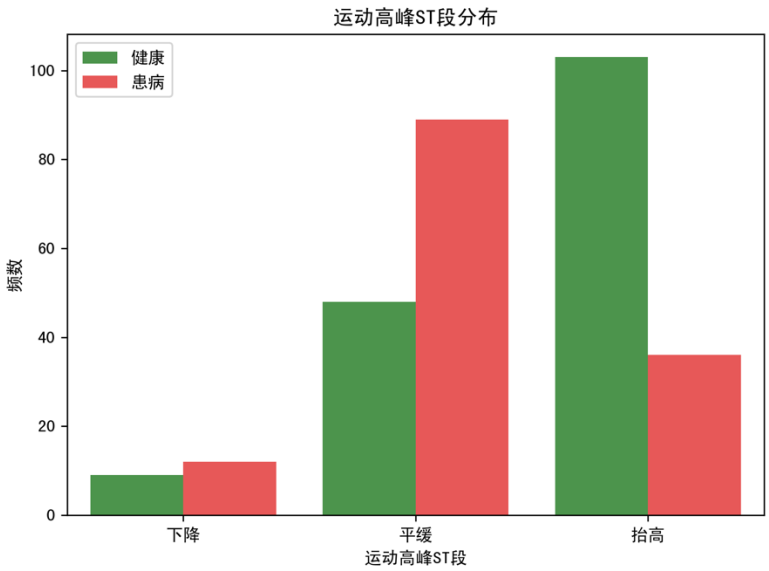
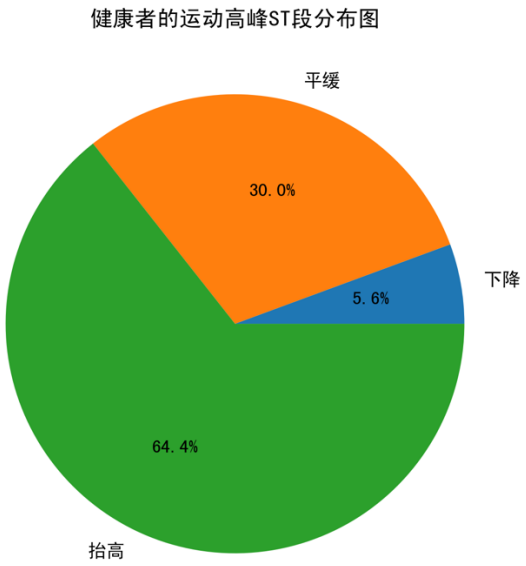


图 17 以年龄为自变量，血清总胆固醇为因变量，不同颜色代表不同诊断结果，通过观察散点图中患病点的分布可以发现，黄色点多集中在偏右和偏上的地方。因此可以推测，年龄集中在 50-70 岁，血清总胆固醇含量越高，患病的风险越大。

图 18，图 19，图 20——运动高峰 ST 段分布





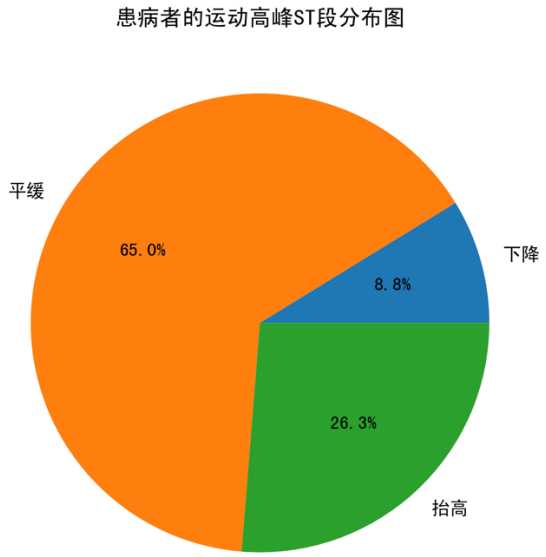
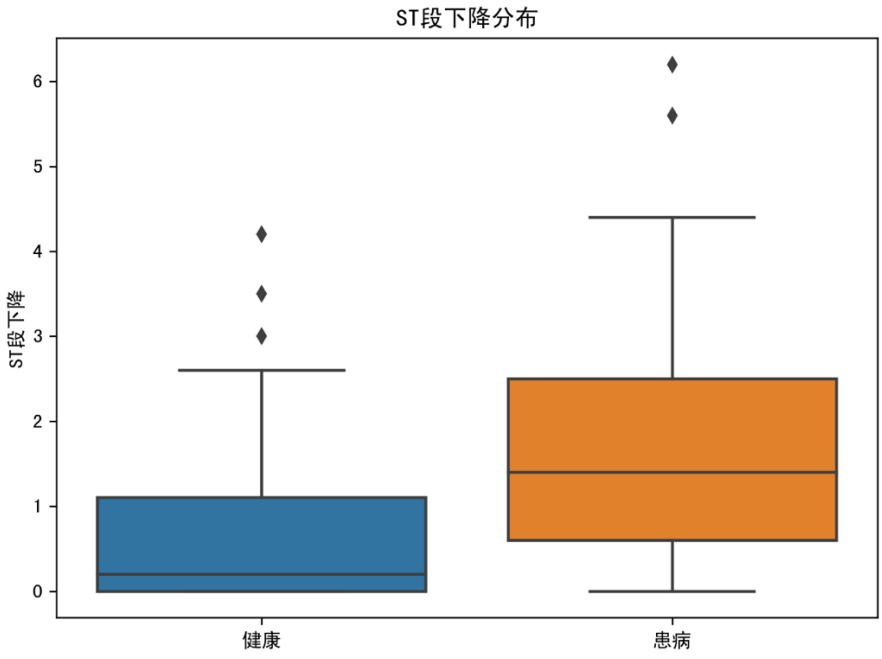


图 18：以运动高峰 ST 段类型作为 x 轴，以出现的频数作为 y 轴，以健康/患病分开，绘制频数分布直方图

图 19、图 20:在健康者、患病者群体中，使用饼状图表示运动高峰 ST 段各类型所占比例

从上述图表中可以明显看出，运动高峰 ST 段平缓的人更可能患有心脏病，运动高峰 ST 段抬高的人更可能是健康的。

图 21，图 22——ST 段下降



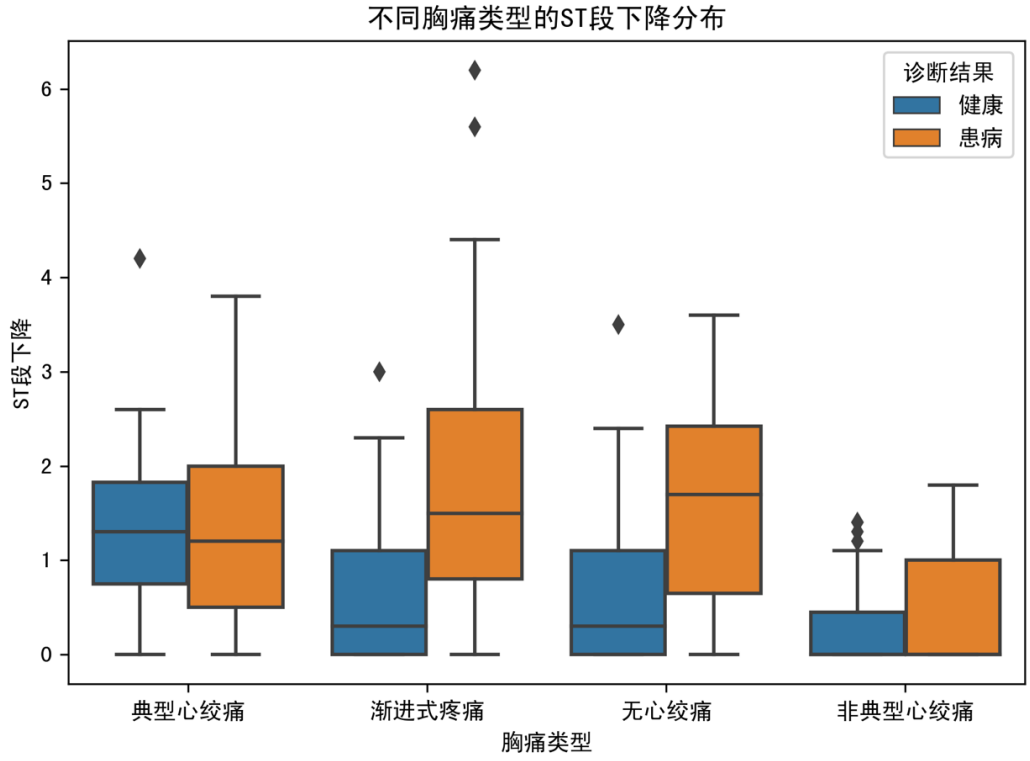


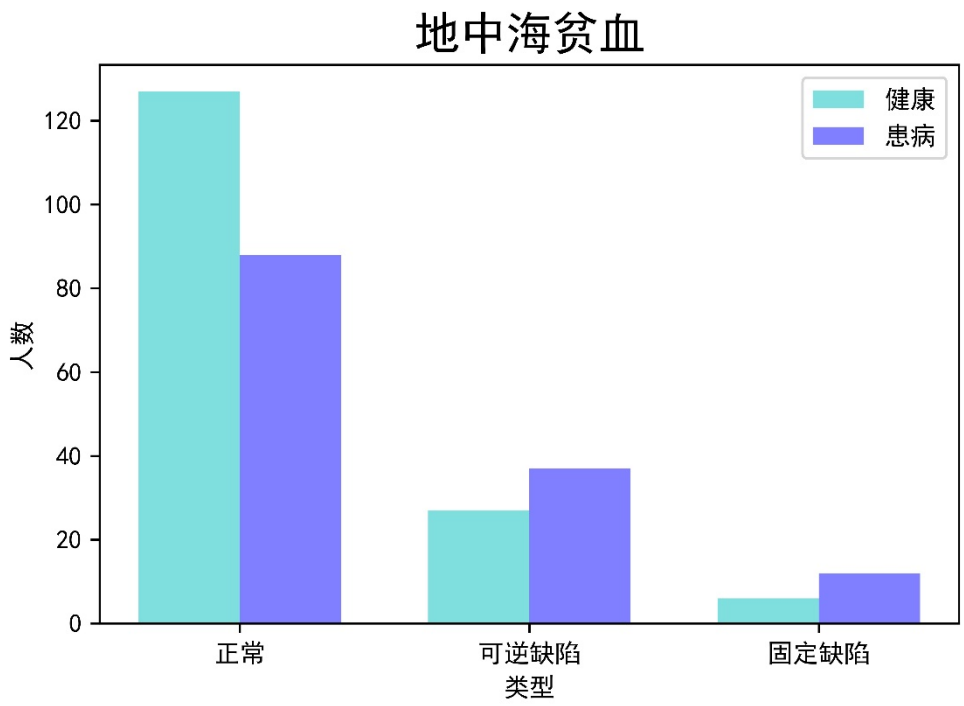
图 21:在健康者和患病者人群中，ST 段下降的分布情况，分别用箱线图表示。

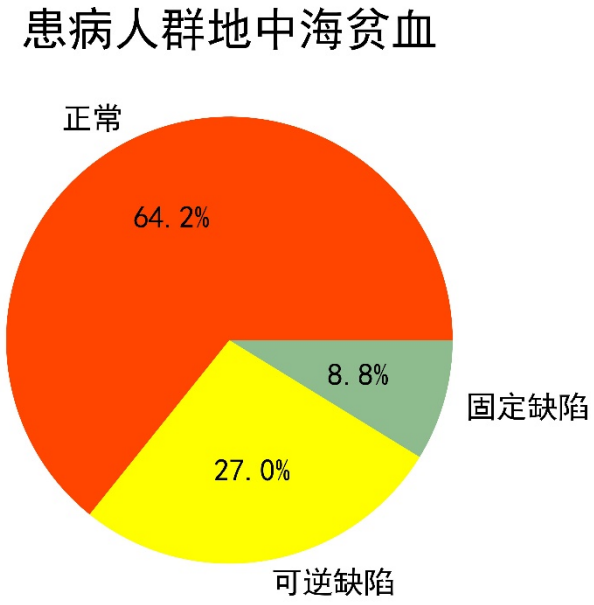
图 22:以胸痛类型为 x 轴，绘制 ST 段下降分布情况的箱线图，并以健康/患病分开。

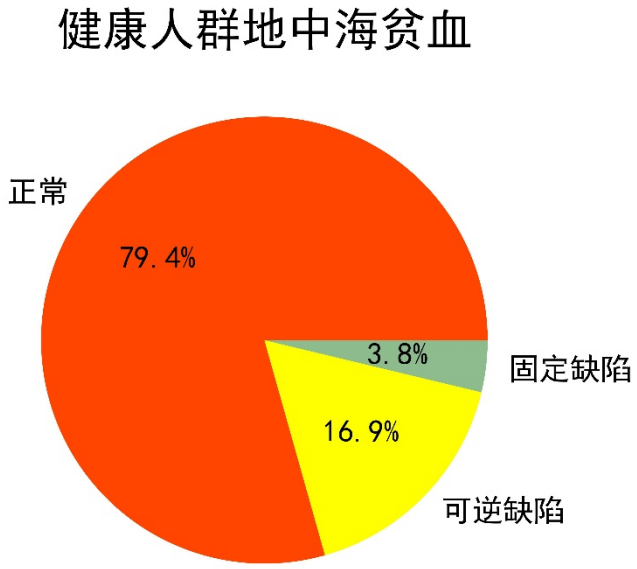
从图 21 中可以看出，患病者的 ST 段下降程度显著更大。而且，对于不同胸痛类型，有如下结论：有典型心绞痛症状者 ST 段下降程度在健康/患病者之间差距不大，其他情况下患病者的 ST 段下降程度更大；而且非典型心绞痛症状的病人 ST 段下降程度普遍偏低。

也就是说，患者有典型心绞痛症状时，ST 段下降程度不能作为进一步诊断的依据；对于渐进式疼痛或无心绞痛的患者，若 ST 段下降程度超过 1mm，便可以一定程度地指示心脏病；而患者有非典型心绞痛症状时，只要 ST 段下降程度高于 0.5mm，即可一定程度上指示心脏病。

图 23，图 24，图 25，图 26——地中海贫血







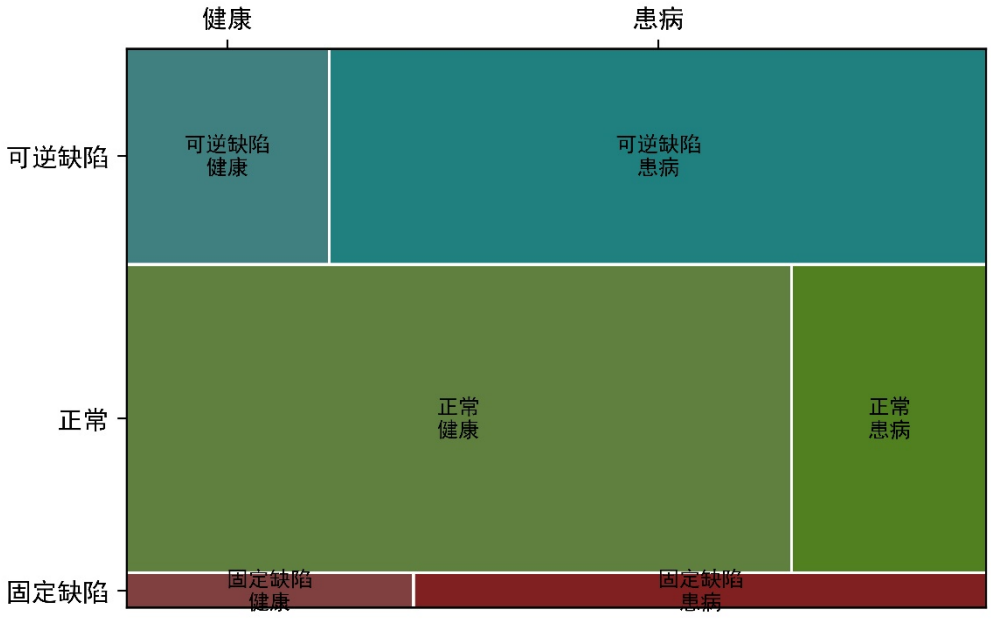


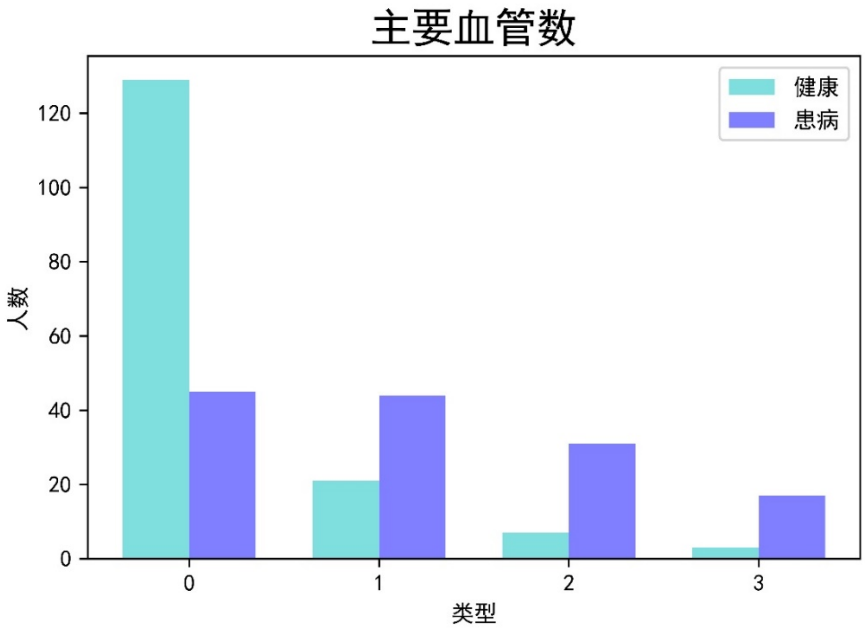
图 23：以地中海贫血病类型为自变量，健康/患病人数为因变量，以颜色区分类别绘制条形图。可以看到，只有在未患地中海贫血症的人群中，心脏健康的人数才大于患病人数，说明是否患有地中海贫血病与心脏健康与否具有一定的联系。

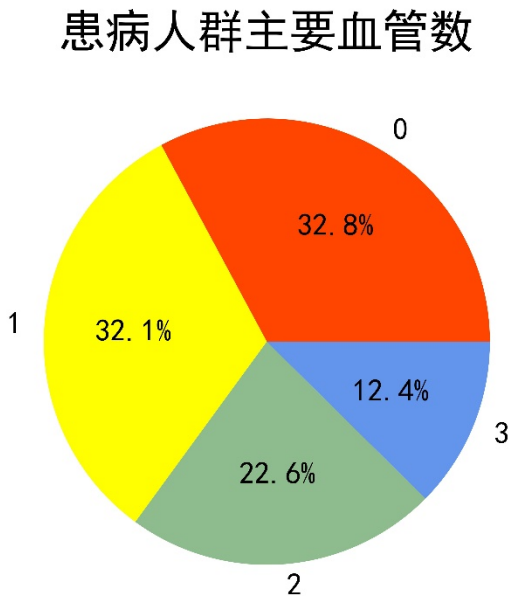
图 24/图 25：以地中海贫血病类型为自变量，人数占比为因变量，不同颜色代表不同的地中海贫血病类型，分别对健康者和患者数据绘制出饼图。两图对比也可

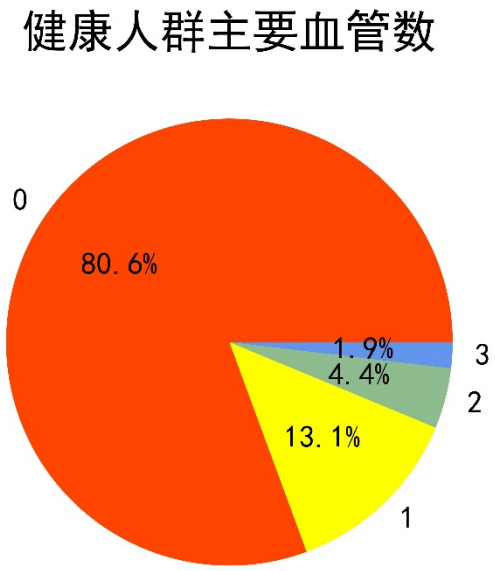
以明显看出，健康人群中地中海贫血病表征正常的比例显著大于患有心脏病人群中的，同样印证了图 23 的分析结论。

图 26：图中每一块面积代表健康或者患病人群中地中海贫血病情况，由此图进一步验证：地中海贫血病表征正常的人群更不易患有心脏病，而两种不同的地中海贫血病对心脏病并无显著影响。

图 27，图 28，图 29，图 30——主要血管数







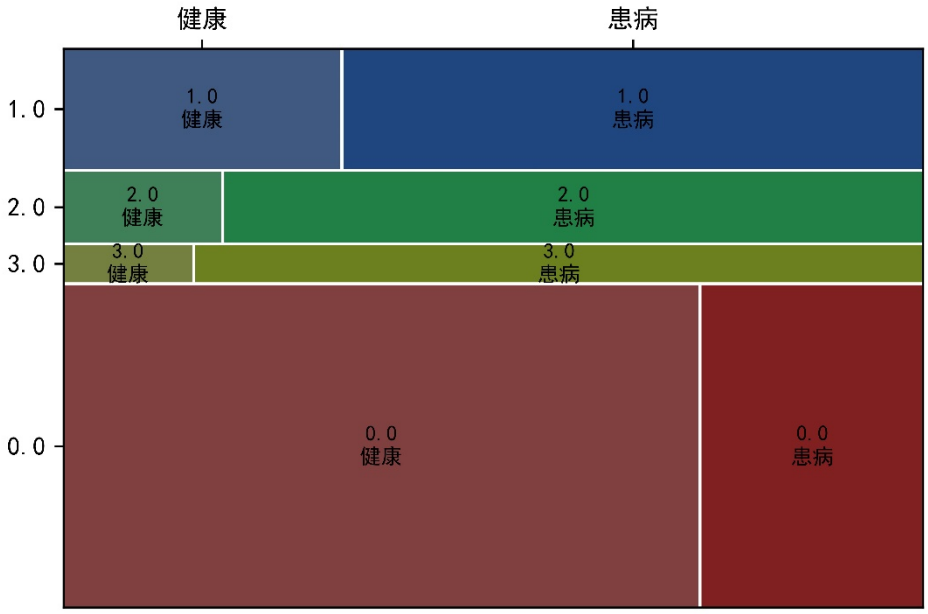


图 27：以主要血管数为自变量，健康/患病人数为因变量，以颜色区分类别绘制条形图。可以看到，在主要血管数为 0 的人群中，心脏健康的人量远高于患心脏病人量，而其他情况均以患心脏病为多数，说明主要血管数量与心脏健康呈现负相关关系。

图 28/图 29：以主要血管数为自变量，人数占比为因变量，不同颜色代表不同的

地中海贫血病类型，分别对健康者和患者数据绘制出饼图。两图对比也可以明显看出，健康人群中主要血管数为 0 占了绝大多数，而患心脏病人群中则呈现出相对均匀的分布。这一方面印证图 27 的分析结论，还说明了在群体基数差异较大时，饼图并不能准确反映情况。

图 30：图中每一块面积代表健康或者患病人群中主要血管数情况，由此图我们可以进一步验证：主要血管数为 0 的人群更不易患有心脏病，但主要血管数进一步增加对患心脏病几率无显著影响。

五、模型选择

本次研究的问题是一个二分类问题，故选择了 Logistic 回归、决策树、随机森林、支持向量机、神经网络五种机器学习模型作为候选。

我们使用 10 折交叉验证，来获取模型在整个数据集上的预测准确率，作为其泛化性能的估计。在交叉验证中，由于每个样本都会出现在测试集中一次，因此我们取这一次的预测类别作为该样本在交叉验证中的预测类别。最后，我们将所有样本的预测类别和真实类别进行比较，得到预测的准确率和混淆矩阵，从而进一步得到查全率、查准率、F1 分数等。

我们使用准确率作为模型选择的依据。交叉验证中准确率越高的模型，被视为在该数据集上具有越好的泛化性能。

同时，我们进行了一定的调参。为降低复杂性，我们没有逐个调整各模型的所有参数，而是选取了一个代表性的参数，并给它一个有限的取值范围，在此范围内寻找使得准确率最高的参数的值。例如，对于决策树模型，我们对 max_depth（树最大深度）参数进行了调参，实验了 1、2、3、5、10、无上限六种取值。

最终，经过多次交叉验证，我们发现 Logistic 回归模型、树最大深度为 2 的随机森林模型、树最大深度为 3 的随机森林模型三者的准确率在所有模型中最高、最稳定。因此，我们认为这三个模型在该数据集上的泛化性能最强，最适合用于本次研究。

由于模型训练中的随机性，每次交叉验证的结果有细微差异。为了复现的方便，这里展示的是所有模型的 random_state 设为 0 时的运行结果：

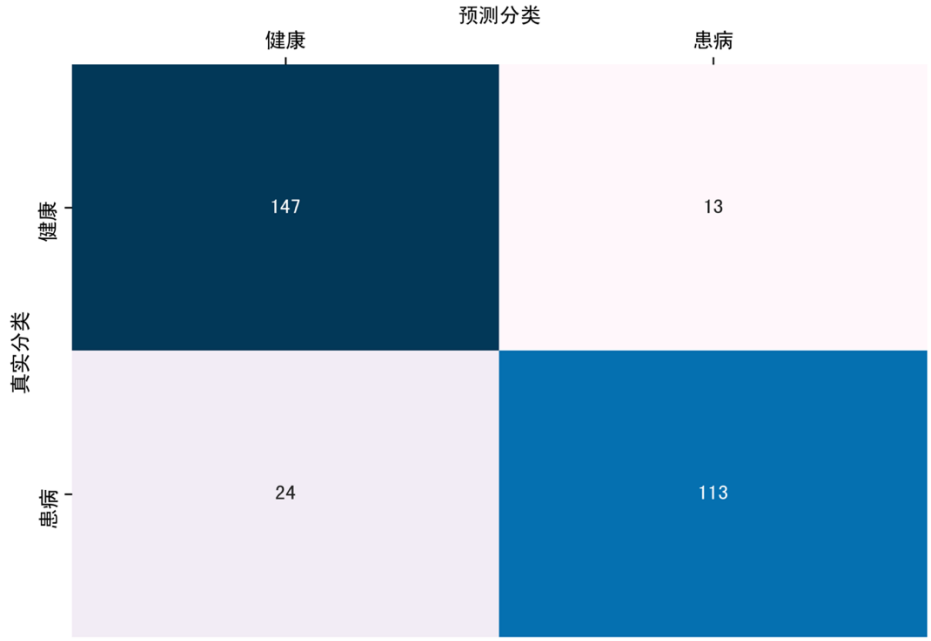
```
Logistic回归: 84.18%
决策树（最大深度=1）：74.07%
决策树（最大深度=2）：72.39%
决策树（最大深度=3）：77.78%
决策树（最大深度=5）：74.41%
决策树（最大深度=10）：74.07%
决策树（最大深度=None）：74.07%
随机森林（最大深度=1）：83.16%
随机森林（最大深度=2）：84.51%
随机森林（最大深度=3）：84.18%
随机森林（最大深度=5）：83.84%
随机森林（最大深度=10）：82.83%
随机森林（最大深度=None）：84.51%
支持向量机（核函数=linear）：81.82%
支持向量机（核函数=poly）：82.83%
支持向量机（核函数=rbf）：81.48%
支持向量机（核函数=sigmoid）：82.49%
神经网络（隐层数量=1，每个隐层的节点数=50）：78.45%
神经网络（隐层数量=2，每个隐层的节点数=25）：77.44%
神经网络（隐层数量=3，每个隐层的节点数=20）：80.81%
神经网络（隐层数量=5，每个隐层的节点数=10）：76.43%
```

可以看到，决策树和神经网络的准确率总体偏低，不超过 80%；支持向量机的准确率在 82%上下；Logistic 回归和随机森林的表现最好。Logistic 回归的准确率稳定在 84%以上。在所有随机森林模型中，树最大深度为 2 和 3 的模型准确率最高，可以达到 84%以上，其他取值下的准确率在 83%上下。（此次运行结果中树最大深度无上限的随机森林模型同样达到了 84.51%的准确率，但这并不常见，在 random_state 取其他值时它便会降到 83%以下）这个结果验证了我们刚才的结论。

六、模型预测

我们选择 Logistic 回归模型、树最大深度为 2 的随机森林模型、树最大深度为 3 的随机森林模型三个模型，在整个数据集上训练，并输出准确率、混淆矩阵、查准率、查全率、F1 分数、ROC 曲线和相应的 AUC 等一系列性能评估指标。

以 Logistic 回归模型为例，它在整个数据集上训练后，结果如下：

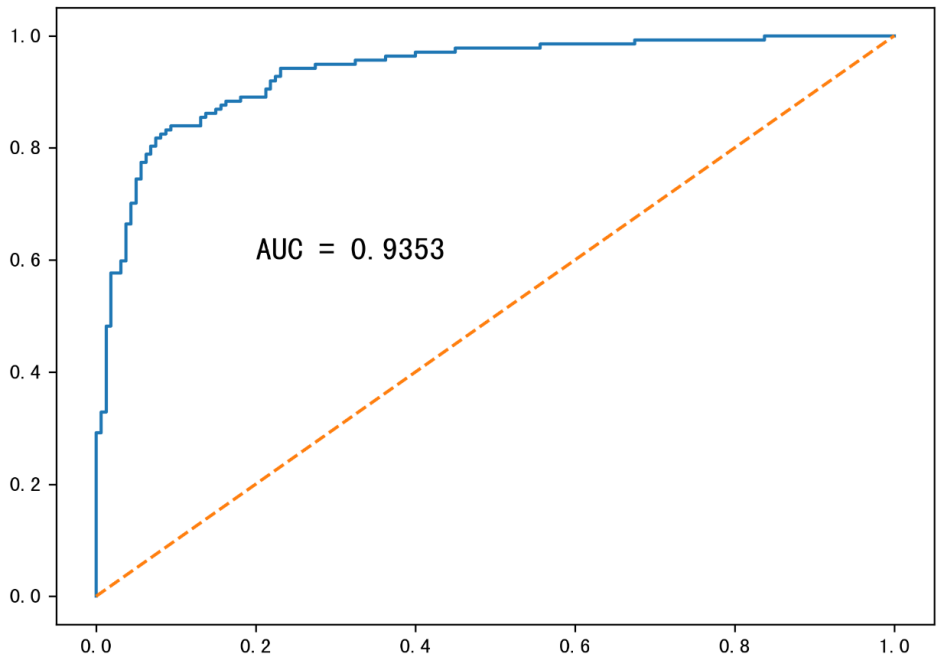


Logistic回归混淆矩阵

Logistic回归

	precision	recall	f1-score	support
健康	0.8596	0.9187	0.8882	160
患病	0.8968	0.8248	0.8593	137
accuracy			0.8754	297
macro avg	0.8782	0.8718	0.8738	297
weighted avg	0.8768	0.8754	0.8749	297

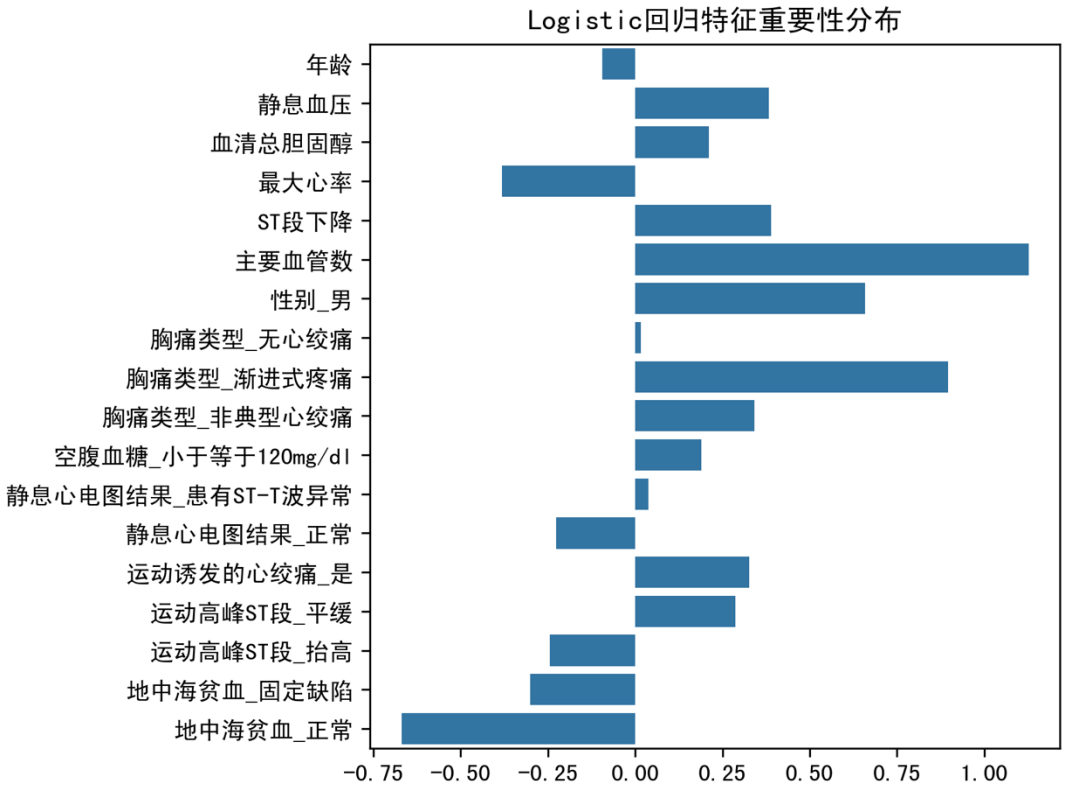
Logistic回归ROC曲线



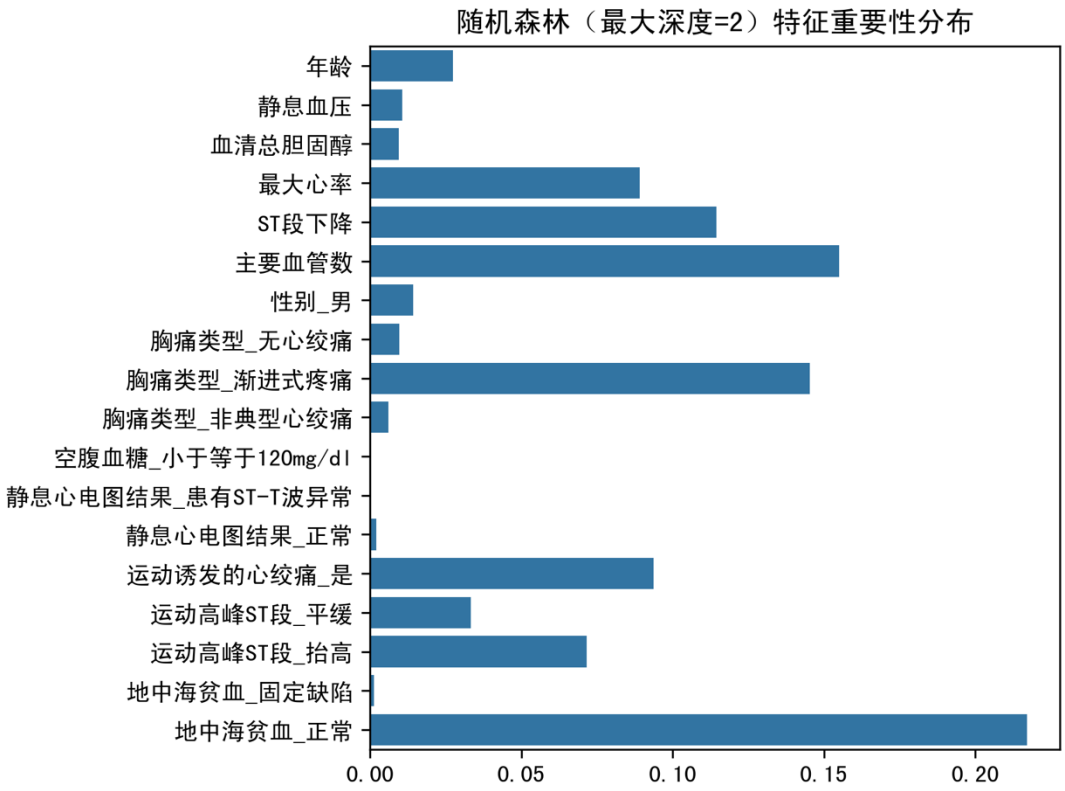
可以看到，模型的预测准确率为 87.54%。对于健康或患病类别，模型的 F1 分数都在 0.85 以上。ROC 曲线的 AUC 达到 0.9353。

另两个模型也能取得类似的令人满意的结果。故这三个模型都可以用于诊断新出现的病人。只要一位病人接受医院的检查，获得他身上本次数据集全部 13 个特征的取值（或者大部分特征的取值，缺失值可以通过插值法填充），模型便能够以很高的准确率诊断出他是否患病。

除此之外，通过训练后的模型的各项系数，我们可以了解哪些特征对诊断结果的影响最大。对于 Logistic 回归模型，其在某一个特征上的系数绝对值越大，说明相关性越强；正负性则可以标示正相关或负相关。如图：



而对于随机森林模型，则用基尼重要性来衡量一个特征的重要性（以这个特征划分数据集时数据集总基尼系数的下降程度）。以树最大深度为 2 的随机森林模型为例：



从特征重要性分布中我们可以看到，主要血管数、胸痛类型、地中海贫血是对诊断结果影响最大的三个特征；ST 段下降、最大心率、运动高峰 ST 段、运动诱发的心绞痛次之；空腹血糖和静息心电图结果对诊断结果几乎无影响。

七、结论与建议

1. 结论：

在第一步的描述性分析中，我们根据数据性质，选用多种图表，对所选数据集中的 13 个自变量进行了可视化描述，该部分的研究帮助我们建立起对于数据集的主观认知，得到一个初步结论：数据集中各个自变量均在不同程度上与心脏病患病与否存在着相关性，对于主要血管数、地中海贫血病等变量的描述甚至直观表现出了相当显著的相关性。

在第二步的建模分析中，我们选取多种常见模型用于目标数据集，经过验证筛选，最终挑选出效果最优的三个模型，它们在数据集上的预测效率达到了类似的高水平，预测准确率均接近九成。根据模型表现，最终我们得到以下两个结论：

A. 本次选取的 UCI 机器学习库中的克利夫兰心脏病数据集中包含了大量有效的心脏病的成因与表征信息，在现代医学的心脏病研究中具有极高的价值，

B. 通过收集分析基础体征信息，可以做到有效预测心脏病患病概率，这证明了在现行医疗体系中建立起一套较为完善的心脏病预测防治机制是一项具有相当可行性的工作。

2. 建议：

鉴于当今心脏病患病率及死亡率仍处于上升阶段的严峻形势，再结合本次研究的结论，我们认为，有可能也有必要，将基层体检与医疗大数据技术相结合，全面建立起心脏病的预测防治体系，这不仅关乎国计民生，更具有重大的战略性意义。

同时，从建模分析中的自变量重要性分布部分，我们还可以直观看出，各个自变量在预测效率上并不相同，但无论对于哪种模型，主要血管数、胸痛类型、地中海贫血重要性均最强，ST 段下降、最大心率、运动高峰 ST 段、运动诱发的心绞痛则次之，而空腹血糖和静息心电图结果对分析几乎无贡献。

于是，从可行性与效率两个方面考虑，若要利用数据大规模地预测防治心脏病，就有必要对预测指标进行进一步优化，将主要血管数、胸痛类型、地中海贫血，ST 段下降、最大心率、运动高峰 ST 段、运动诱发的心绞痛此七类特征纳入常规体检项目，并利用健康医疗大数据对这些标记的数据特征进行实时预测，更新患心脏病概率，从而建立起有效的心脏病预测防治体系。