企业智能系统项目实作报告

题目：运用随机森林方法于心脏病预测领域分析

学生姓名：吕循新、叶兴炼

学　　号：201843307123、201843302129

系　　别：计算机科学与技术系（跨境电商）

目录

一、绪论

1.1 研究背景

随着数据量以及计算机性能的不断提升，机器学习技术正逐渐渗透于各行各业中。计算机视觉、自然语言处理、机器人等领域基本上已经被机器学习算法垄断，正逐步向教育、银行、医疗等传统行业扩张。而目前在医疗行业对机器学习的应用也逐步成熟，通过机器学习，医疗服务提供商可以通过机器学习对患者的诊断和治疗选择做出更好的决策，从而导致医疗服务的整体改善。

以前，由于没有可用的技术或工具，医疗保健专业人员收集和分析大量数据以进行有效的预测和治疗是具有挑战性的。现在，随着机器学习的发展，Hadoop 等大数据技术已经足够成熟以适应大规模应用，这些工作相对来说就变得比较容易。实际上，54％的企业正在使用或将 Hadoop 作为大数据处理工具，以获得有关医疗保健的重要见解。94％的 Hadoop 用户对以前认为不可能的庞大数据进行分析。

机器学习算法也可以帮助提供关于患者疾病、实验室检查结果、血压、家族史、临床试验数据等方面的重要统计数据、实时数据和高级分析。

由于医疗保健产生大量数据，所面临的挑战是收集这些数据并将其有效地用于分析、预测和治疗。而心脏病是人类健康的头号杀手。全世界 1/3 的人口死亡是心脏病引起的。而我国，每年有几十万人死

于心脏病。如果可以通过提取人体相关的体测指标，通过数据挖掘方式及机器学习来分析不同特征对于心脏病的影响，将对预防心脏病起到至关重要的作用。

1.2 研究动机

本次分析目的是通过大数据分析与建模，通过提取人体相关的体测指标找出与心脏病相关的重要变量，通过数据挖掘方式及机器学习来分析不同特征对于心脏病的影响，将对预防心脏病起到至关重要的作用。希望通过该模型可以让普通民众通过该模型可以预测到自己是否有患有心脏病的潜在风险，尽早发现尽早治疗，为国内心脏病研究提供一定的贡献。

1.3 研究目的

为了达到提升心脏病预测准确率的目的,研究将分为特征工程中的数据预处理和随机森林两部分展开讨论.首先,根据 UCI 克利兰夫医学研究中心的心脏病数据集,通过获得患者某些心血管疾病的重要指标,用数据预处理技术改善数据质量,使模型结构在预测方面和特征的关联性方面均有加强.在探讨不同类型的参数对网络结构的影响,并且通过 GridSearchCV 构建并优化了心脏病预测模型,为心脏病疾病的防治提供参考.经过与传统心脏病预测方法比较,证明了研究提出的基于随机森林的心脏病预测方法分类更加准确.辅助医生判断患者是否有心脏病,使计算机不仅在医疗上用于检测,更能应用于民众的自行预测,为人民对自己的身体情况提供参考，同时较好地避免人为导致的误差。

二、文献探讨

目前国外诸多学者对心脏病发病预警模型进行了研究[i]。2009 年，Tan K C,Teoh E J[ii]提取加州欧文分校机器学习数据库心脏病数据集，在 LIB 支持向量机和 Weka 上实现，得到 84.07%的预测准确率。Chaurasia、Pal[iii]在 2013 年使用朴素贝叶斯、J48、引导聚集算法对 UCI 数据集中的 11 个特征项进行预测，获得结果显示朴素贝叶斯准确率为 82.31%,J48 准确率为 84.31%,引导聚类算法准确率为 85.03%。Parthiban、Srivat - sa[iv]2012 年利用来自印度金奈某研究所的心脏病数据集使用 Weka 平台实现朴素贝叶斯及支持向量机诊断心脏病患病率，分别得到准确率 74.00%、94.60%。2015 年 Vem bandasamy 等人[v]使用朴素贝叶斯算法对印度金奈某研究所的心脏病数据集进行分类预测，得到 86.42%预测准确率。而在国内叶苏婷[vi]等人使用决策树的方法来预测心脏病，得到了 81.81%的准确率机器学习算法涵盖广泛，在模型研究时，特征变量，算法的选择不同，均会导致预测准确率差异。

本研究采用随机森林的算法对 UCI 克利兰夫医学研究中心的心脏病数据集构建研究模型。

三、研究方法

3.1 随机森林

3.1.1 原理说明

随机森林是一种有监督学习算法。它创建了一个森林，并使它拥有某种方式随机性。 所构建的“森林”是决策树的集成，大部分时候都是采用“bagging”方法训练的。 bagging 方法，即 bootstrap aggregating，采用的是随机有放回的选择训练数据然后构造分类器，最后组合学习到的模型来增加整体的效果。

随机森林建立了多个决策树，并将它们合并在一起以获得更准确和稳定的预测。随机森林的一大优势在于它既可用于分类，也可用于回归问题，这两类问题恰好构成了当前的大多数机器学习系统所需要面对的。

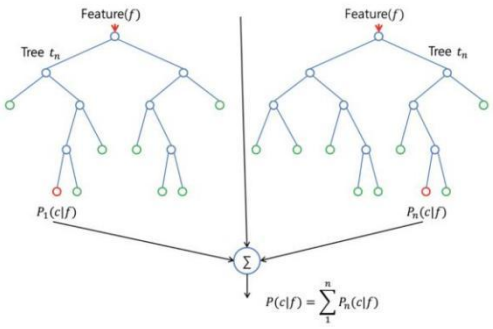

图 3-1-1 两棵树的随机森林

随机森林分类器使用所有的决策树分类器以及 bagging 分类器的超参数来控制整体结构。与其先构建 bagging 分类器，并将其传递给决策树分类器，使用者可以直接使用随机森林分类器类，这样对于决策树而言，更加方便和优化。要注意的是，回归问题同样有一个随机森林回归器与之相对应。

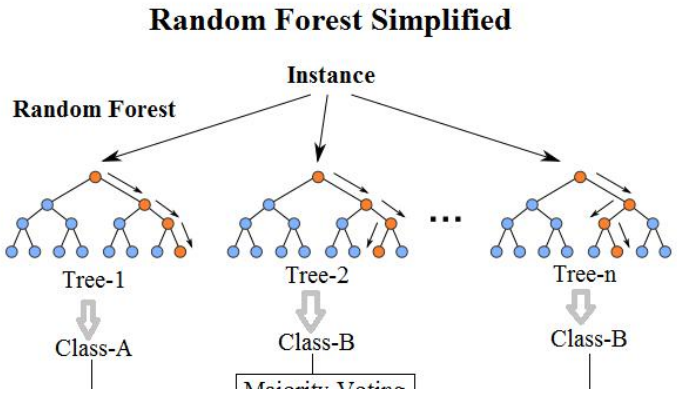

图 3-1-2 随机森林图示

随机森林算法中树的增长会给模型带来额外的随机性。与决策树不同的是，每个节点被分割成最小化误差的最佳特征，在随机森林中我们选择随机选择的特征来构建最佳分割。因此，当您在随机森林中，仅考虑用于分割节点的随机子集，甚至可以通过在每个特征上使用随机阈值来使树更加随机，而不是如正常的决策树一样搜索最佳阈值。

这个过程产生了广泛的多样性，通常可以得到更好的模型。

3.1.2 本案例使用随机森林的优点

随机森林模型可以很容易地测量每个特征对预测的相对重要性。并且 Sklearn 为此提供了一个很好的工具，它通过查看使用该特征减少了森林中所有树多少的不纯度，来衡量特征的重要性。它在训练后自动计算每个特征的得分，并对结果进行标准化，以使所有特征的重要性总和等于 1。

通过查看特征的重要性，我们可以知道哪些特征对预测过程没有足够贡献或没有贡献，从而决定是否丢弃它们这会**比决策树更好找到**

与本案例相关的重要特征。另一个区别是“深度”决策树往往会遭遇过拟合问题。而随机森林则可以通过创建随机的特征子集并使用这些子集构建较小的树，随后组成子树，这种方法可以**防止大部分情况的过拟合**。

所以本次研究选用的模型为随机森林模型，总结选用这个模型的目的是：

1）每棵树随机选择样本并随机选择特征，使得具有很好的抗噪能力，性能稳定；

2） 因为本次分析的特征较多而且大部分都是类别型数据，需要进行特征工程，而随机森林能处理很高维度的数据，并且不用做特征选择；

3） 实现比较简单。

3.2 整体实验流程

3.2.1 整体实验图示

图 3-2-1 数据挖掘图示

3.2.2 变量检视

研究采用来源 UCI 机器学习知识库的心脏病数据集，共有 303 个样本和 76 个特征，基于该数据集选用心脏病致病原因的 14 个特征构成特征子集，如下图所示。

序号	编号	特征字段名	基本描述
1	#3	年龄(age)	检测对象年龄
2	#4	性别(sex)	检测对象性别，男性=1、女性=0
3	#9	胸部疼痛类型(cp)	4个值，典型心绞痛=1、非典型心绞痛=2、非心绞痛性疼痛=3、无症状=4
4	#10	静息血压(trestbps)	静息血压值，以医院毫米汞柱为单位
5	#12	血清胆固醇(chol)	血清胆固醇浓度(mg/dl，毫克每分升)
6	#16	空腹血糖值(fbs)	空腹血糖量大于120mg/dl,是=1，否=0
7	#19	静息心电图结果(restecg)	3个值，正常=0、ST-T波异常(T波倒置和/或ST段抬高或压低>0.05mV)=1、Estes标准下存在心室肥厚=2
8	#32	最大心率(thalach)	达到的最大心率值
9	#38	运动诱发心绞痛(exang)	是=1；否=0
10	#40	ST段压值(oldpeak)	运动引起的相对于休息的ST段压低
11	#41	ST段倾斜度(slope)	峰值ST段斜率，上升=1、平=2、下降=3
12	#44	可检测血管数目(ca)	主要血管(0-3)的数量用荧光法着色
13	#51	缺陷种类(thal)	正常=3、固定缺陷=6、可逆缺陷=7
14	#58	是否患病(target)	是=1；否=0(血管造影疾病状态)

图 3-2-2 特征变量说明

3.2.3 数据预处理

在数据预处理的过程中我们需要根据每个字段的含义将字符型转为数值。

1）二值类的数据

二值类的比较容易转换,如 sex 字段有两种表现形式 female 和 male，我们可以将 female 表示成 0，把 male 表示成 1。

2）多值类的数据

比如 cp 字段，表示胸部的疼痛感，我们可以通过疼痛的由轻到重映射成 0~3 的数值。

3.2.4 特征工程

特征工程主要是包括特征的衍生、尺度变化等。本例中有两个组件负责特征工程的部分。

1）过滤式特征选择

主要是通过这个组件判断每个特征对于结果的影响，通过信息熵和基尼系数来表示，可以通过查看 SQL 中的评估报告来显示最终的结果。

图 3-3-1 熵和基尼系数

2）归一化

因为本次实验选择的是通过逻辑回归二分类来进行模型训练，需要每个特征去除量纲的影响。归一化的作用是将每个特征的数值范围变为 0 到 1 之间。并且归一化有着两大优点：

1. 归一化后加快了梯度下降求最优解的速度；

2. 有可能提高精度。

3.2.5 模型搭建

在 python 中导入 Sklearn 中 train__test__split，将数据集分割为训练集和测试集，设测试集占比 20%，并且搭建随机森林模型。与传统的调参不同，本次研究会用网格搜索来进行高效的参数调优，具体实现方法为 sklearn 里面 GridSearchCV 进行高效调参，GridSearchCV 根据会根据给定的模型自动进行交叉验证，通过调节

每一个参数来跟踪评分结果，实际上，该过程代替了进行参数搜索时的 for 循环过程。最后得出是否患有心脏病的预测模型。

1）最佳随机森林模型参数

经过 GridSearchCV 进行高效调参，后发现最佳模型为 max_depth = 5，min_samples_leaf = 1， min_sample_split = 2， n_estimators = 10， random_state = 5。

```
#最佳模型
model4=RandomForestClassifier( max_depth=5, max_features='sqrt',
                        min_samples_leaf=1, min_samples_split=2,
                         n_estimators=10, random_state = 5, oob_score=True)
model4.fit(train_X,train_y)
```

图 3-2-4 最佳随机森林模型

2）决策树可视化

任意在随机森林中挑选一颗决策树进行可视化

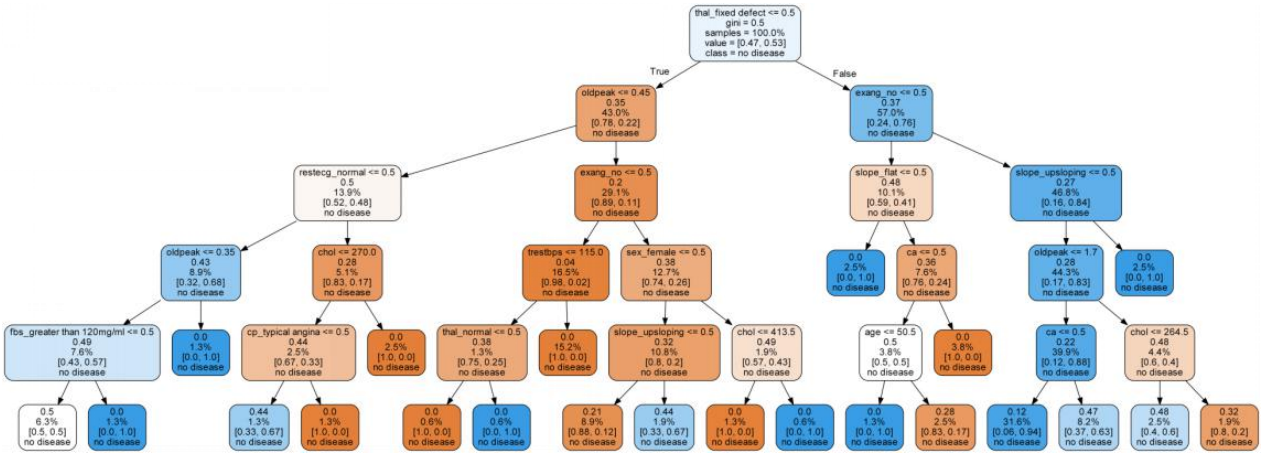

图 3-2-5 决策树可视化

1.查看特征中的权重以及对特征重要性进行排序

```
重要性特征排序
feature ca (0.115223)
feature thal_fixed defect (0.113785)
feature cp_typical angina (0.106293)
feature oldpeak (0.090147)
feature thalach (0.085871)
feature thal_reversable defect (0.069922)
feature exang_yes (0.052044)
feature age (0.050104)
feature chol (0.047142)
feature trestbps (0.045093)
feature slope_downsloping (0.039228)
feature exang_no (0.032542)
feature slope_flat (0.026342)
feature sex_male (0.024862)
feature sex_female (0.019857)
feature cp_non-anginal pain (0.017827)
feature cp_asymptomatic (0.013508)
feature restecg_normal (0.011109)
feature restecg_ST-T wave abnormality (0.010026)
feature fbs_greater than 120mg/ml (0.006396)
feature thal_normal (0.006350)
feature cp_atypical angina (0.006234)
feature fbs_lower than 120mg/ml (0.004978)
feature slope_upsloping (0.003571)
feature thal_unknown (0.000980)
feature restecg_left ventricular hypertrophy (0.000564)
```

Weight	Feature
0.3963	thal_fixed defect
0.1359	oldpeak
0.1184	exang_no
0.0645	ca
0.0641	slope_flat
0.0579	chol
0.0466	restecg_normal
0.0394	age
0.0214	slope_upsloping
0.0177	thal_normal
0.0158	cp_typical angina
0.0101	fbs_greater than 120mg/ml
0.0064	sex_female
0.0054	trestbps
0	thal_reversable defect
0	thalach
0	slope_downsloping
0	exang_yes
0	thal_unknown
0	restecg_ST-T wave abnormality
	... 6 more ...

图 3-2-6 特征重要性排序

特征权重和特征重要性有利于找到关键并且之后对关键特征进行分析

3）模型预测

1.构建混淆矩阵

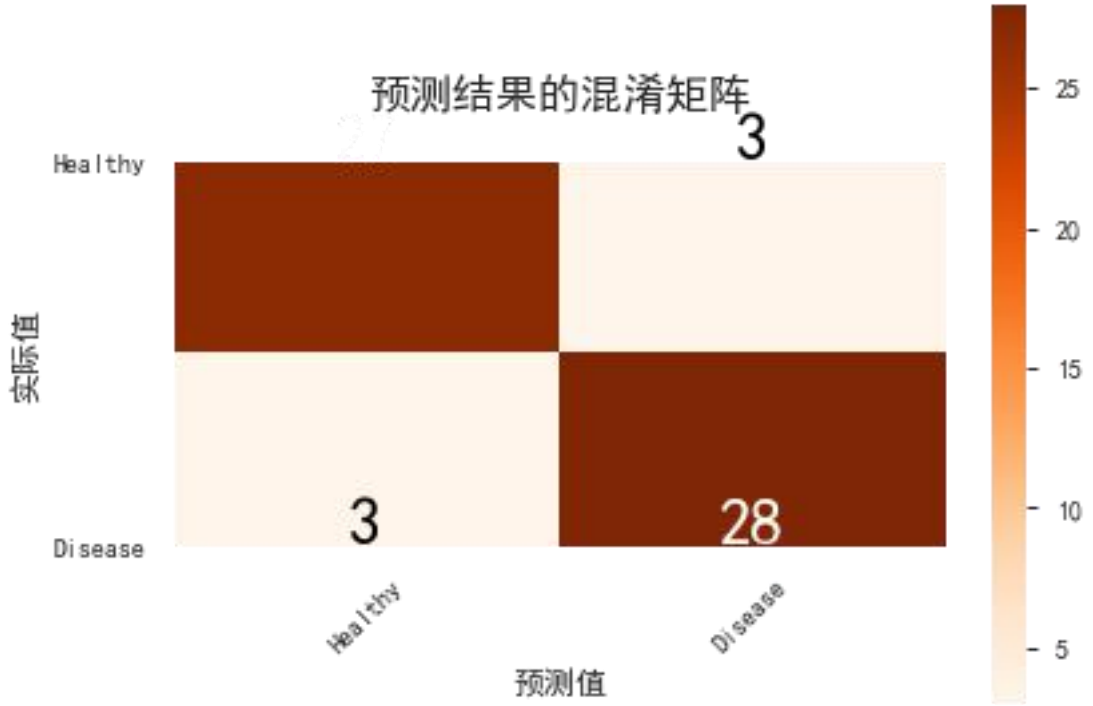

图 3-2-7 混淆矩阵

2.对混淆矩阵进行评分

	precision	recall	f1-score	support
Healthy	0.90	0.90	0.90	30
Disease	0.90	0.90	0.90	31
accuracy			0.90	61
macro avg	0.90	0.90	0.90	61
weighted avg	0.90	0.90	0.90	61

图 3-2-8 混淆评分结果

可见各项指标都是不错的，模型的可靠性很高

3.对模型进行 ROC 的曲线绘制

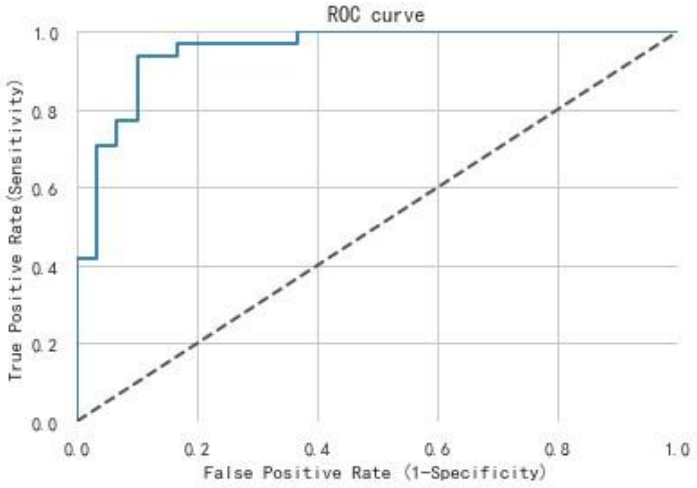

图 3-2-9 ROC 曲线

ROC 曲线指受试者工作特征曲线 / 接收器操作特性曲线 (receiver operating characteristic curve)，是反映敏感性和特异性连续变量的综合指标,是用构图法揭示敏感性和特异性的相互关系，它通过将连续变量设定出多个不同的临界值，从而计算出一系列敏感性和特异性，再以敏感性为纵坐标、（1-特异性）为横坐标绘制成曲线，曲线下面积越大，诊断准确性越高。在 ROC 曲线上，最靠近坐标图左上方的点为敏感性和特异性均较高的临界值。

总结来说 ROC 的曲线越靠近左上角模型越好

四、实验结果与讨论

4.1 实验结果

由上述步骤，逐步得到以下结果通过 GridSearchCV 进行高效调参，得到最佳模型为 max_depth = 5， min_samples_leaf = 1，min_sample_split = 2， n_estimators = 10， random_state = 5。根据模型测试得到了以下评价数值：

该模型的在训练集中的表现如下：

精确率：0.931， 召回率：0.927， F1 值：0.929， R2 值：0.716

该模型的在测试集中的表现如下：

精确率：0.902， 召回率：0.902， F1 值：0.902， R2 值：0.606

各项指标都比较不错，可见没有出现模型过拟合的现象，达到实验结果较理想。

4.2 实验重要特征分析与讨论

1.性别

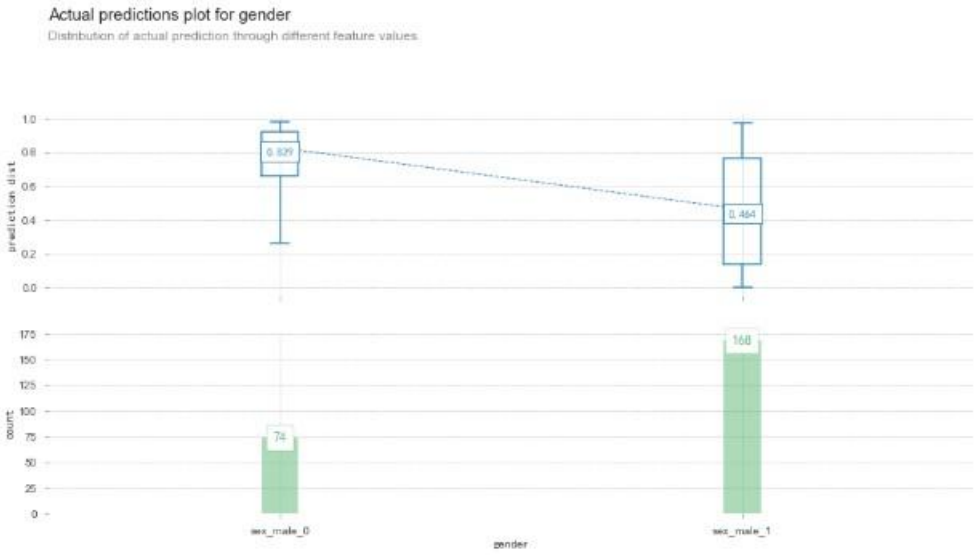

图 4-2-10 性别特征比较

在模型中可看出男性比女性患病概率低

2.大血管的数量

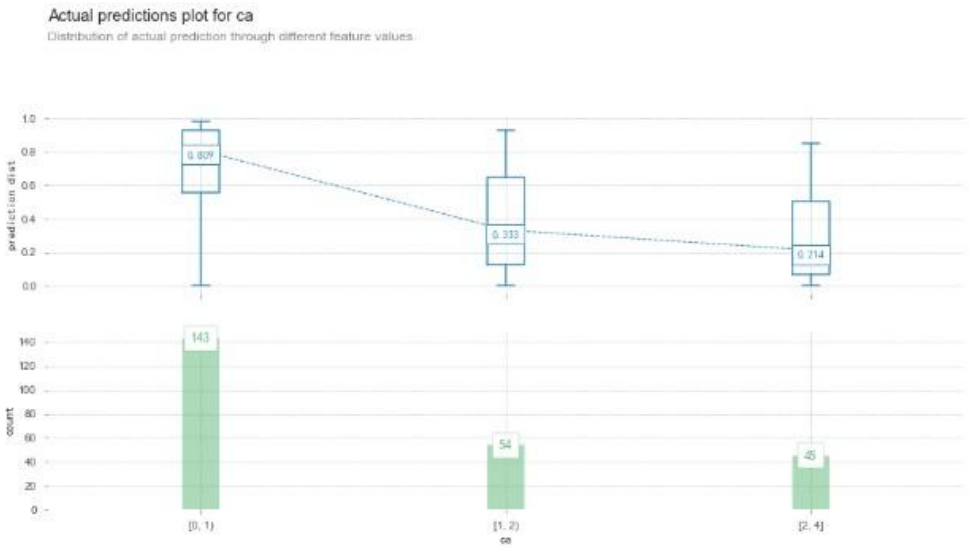

图 4-2-11 血管数特征比较

模型可看出大血管数越多，患病概率越低

3.最大心率

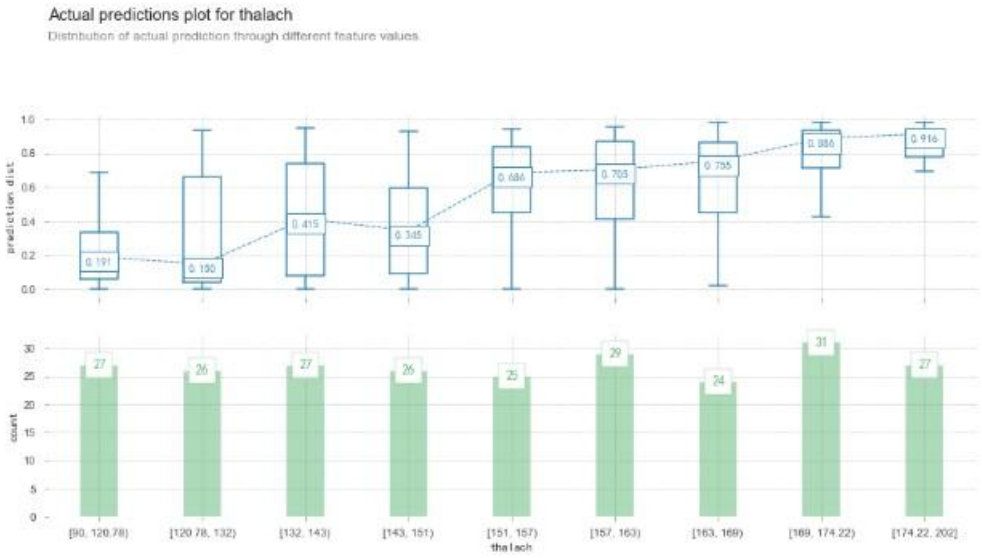

图 4-2-12 最大心率特征比较

模型可看出心率越高患病概率也就越高

4.地中海贫血

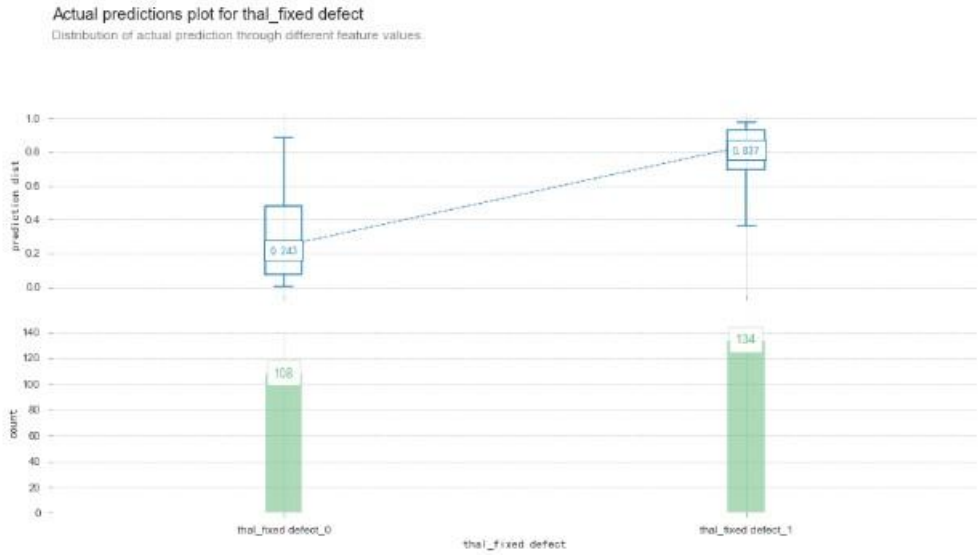

图 4-2-13 地中海贫血特征比较

模型可看出有地中海贫血的患者并且具有固定缺陷的人患病概率会高

5.休息的运动引起的 ST 值变化

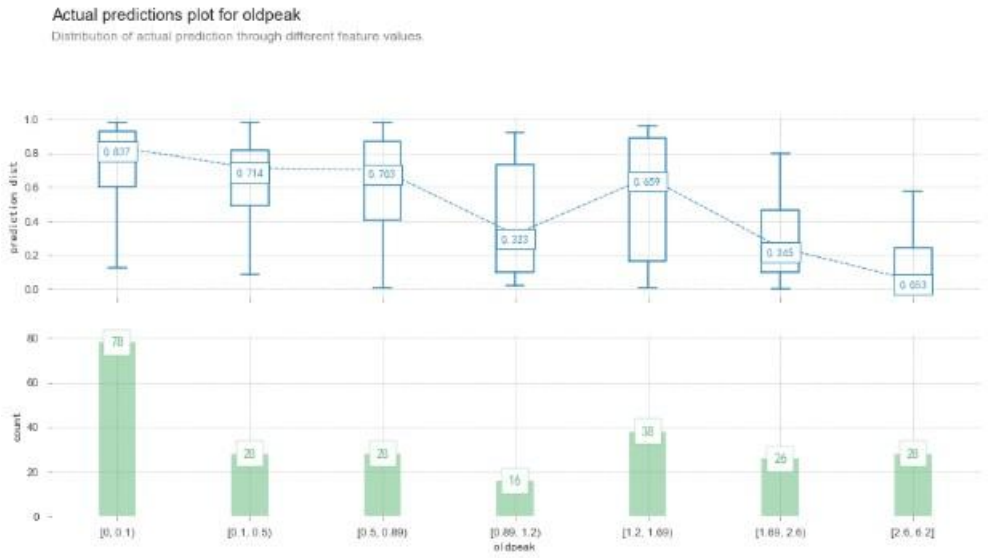

图 4-2-14 ST 值特征比较

模型预测相对于休息的运动引起的 ST 值在 1.8-6.2 这个区间的患病概率是比较低的

6.年龄

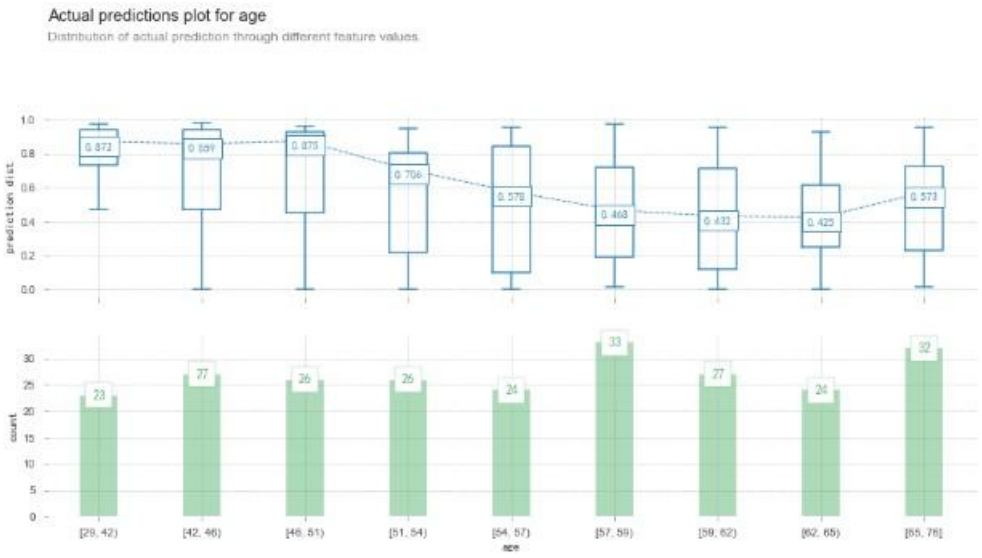

图 4-2-15 年特征比较

模型看出年龄越大患有心脏病概率有降低趋势，与我们固有印象年龄越大越有可能患有心脏病不同

五、结论

研究利用机器学习中随机森林算法，构建心脏病预警模型。为用户构建预警界面，测试患者是否患有心脏病，为临床医生及患者提供预警信息。通过完整的实验流程，决策树算法预测准确率为 0.902，结合其他研究结果，本数据结果相对理想，可较准确的反应患者的患病情况，但预测数据结果仍有小部分的偏差，本模型仍存在改善空间。

本次研究有可能为医生判断一个病人是否患有心脏病提供重要的依据，并且在搭建的模型中心率高低和大血管数量是重要因素这可以为人们预防心脏病和对自己身体出现了异常而及时就医提供了更大的帮助，而女性更需要注意自己的心脏健康，并且中年人是心脏病发

的高危人群，这都可以为人们带来警醒，为减低心脏病发率提供了参考及帮助。

下一步研究重点，将着重改进预警模型的预测准确率，提升模型使用效率，并广泛利用机器学习中算法，对多种疾病进行算法分析、模型构建、预警研究。

---

[i] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic[J]. Journal of Intelligent Learning Systems and Applications, 2017, 9(1):1-16.

[ii] Tan K C, Teoh E J, Yu Q, et al. A hybrid evolutionary algorithm for attribute selection in data mining[J]. Expert Systems With Applications, 2009, 36(4):8616-8630.

[iii] Chaurasia V, Pal S. Data mining approach to detect heart disease[J]. International Journal of Advanced Computer Science and Information Technology, 2013, 2(4):56-66.

[iv] Parthiban G, Srivatsa S K. Applying machine learning methods in diagnosing heart disease for diabetic patients[J]. International Journal of Applied Information Systems, 2012, 3(7):25-30.

[v] [8] Vembandasamy K, Sasipriya R, Deepa E. Heart diseases detection using naïve Bayes algorithm[J]. International Journal of Innovative Science, Engineering and technology, 2015(2):441-444.

[vi] 叶苏婷,潘媛媛,毕迎春.基于决策树算法的心脏病发病预警模型研究[J].电脑知识与技术,2020,16(19):187-189.