

运用随机森林方法于心脏病预测领域分析

吕循新 201843302129 叶兴炼 20184302129

摘要：

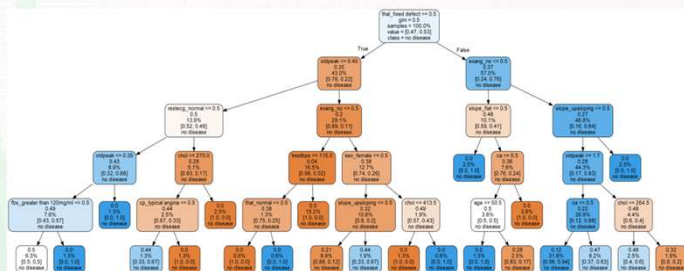
医疗行业对机器学习的应用也逐步成熟，通过机器学习，医疗服务提供商可以通过机器学习对患者的诊断和治疗选择做出更好的决策，从而导致医疗服务的整体改善。由于医疗保健产生大量数据，所面临的挑战是收集这些数据并将其有效地用于分析、预测和治疗。而心脏病是人类健康的头号杀手。全世界1/3的人口死亡是心脏病引起的。而我国，每年有几十万人死于心脏病。通过分析建模，通过提取人体相关的体测指标找出与心脏病相关的重要变量，通过数据挖掘方式及机器学习来分析不同特征对于心脏病的影响，将对预防心脏病起到至关重要的作用，尽早发现尽早治疗。

整体实验流程：



查看模型：

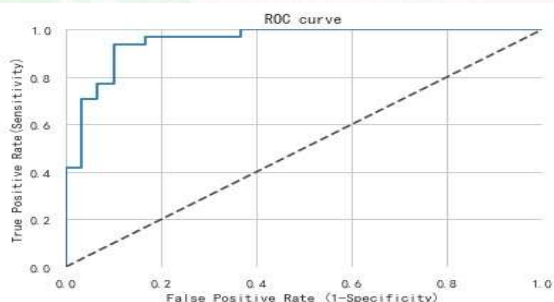
经过GridSearchCV进行高效调参，得到的最佳模型：



可视化：随机森林中一颗决策树进行可视化

模型预测：

模型预测一共使用了三种方法我们挑出期中一种来说明：**ROC曲线**



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。

实验结果：

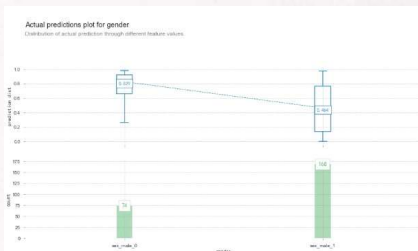
各项指标都比较不错，可见没有出现模型过拟合的现象，达到实验结果较理想。

查看特征中的权重以及对特征重要性进行排序：

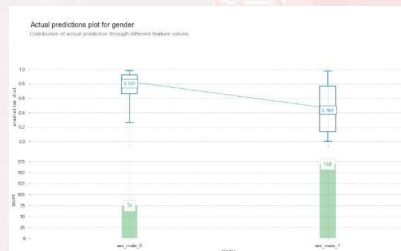
重要性特征排序
feature ca (0.115223)
feature thal_fixed defect (0.113785)
feature cp_typical angina (0.106293)
feature oldpeak (0.090147)
feature thalach (0.085871)
feature thal_reversible defect (0.069922)
feature exang_yes (0.052044)
feature age (0.050104)
feature chol (0.047142)
feature trestbps (0.045093)
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	slope_normal
0.0158	cp_typical angina
0.0101	fbs_greater than 120mg/ml
0.0064	sex_female
0.0054	trestbps
0	thal_reversible defect
0	thalach
0	slope_downsloping
0	exang_yes
0	thal_unknown
0	restecg_ST-T wave abnormality
...	6 more ...

查看重要特征（部分）：



性别：在模型中可看出男性比女性患病概率低



血管数：模型可看出大血管数越多，患病概率越低。

结论：

这里用户构建预警界面，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。搭建的模型中心率高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，为减低心脏病发生率提供了参考及帮助。