上海财经大学

十年内冠心病发病概率探索

探索性数据分析课程论文

范怡雯 2018111201

李晨茜 2018110760

吕郢歌 2018110852

摘 要

冠状动脉疾病又称为缺血性心脏病或简称冠心病，是最常见的心血管疾病。型态包含稳定型心绞痛、非稳定型心绞痛、心肌梗塞和猝死。冠状动脉疾病在公元 2002 年是全球第一大死因，也是人们住院的主要原因之一，2013 年也是全球死因首位，死亡人数自 1990 年 574 万人（12%）攀升至 2013 年 814 万人（16.8%）。研究表明，早期干预危险因素是减少心血管病负担的关键措施，据全球疾病负担系列，改善上述危险因素可减少冠心病发生率 83%~89%，减少冠心病死亡率 78%~85%。因此本报告从预测与预防角度出发，通过构建逻辑回归模型、决策树与随机森林模型以及支持向量机模型并探索模型得到的影响十年内冠心病发病率的重要因素之间的共性与特性，来为存在潜在风险的被调查者提出有效预防建议。

关键词：预防、逻辑回归、决策树与随机森林、支持向量机、模型比较

目录

课题综述

一、选题背景

动脉粥样硬化性疾病是一个逐渐累积的过程，动脉粥样硬化起始于儿童，中年进展最快，老年发病最多。随着生活水平的改善、各种不健康生活方式的流行、肥胖比例越来越大、生活压力的增大，使动脉粥样硬化发展过程缩短了。冠心病患者多存在肥胖（特别是腹型肥胖）、吸烟、大量喝酒、糖尿病、高脂血症、高尿酸血症等危险因素。平时也存在一些不良的生活习惯，比如吸烟、不爱运动、工作压力大、脾气大、饮食不健康等。因此，从预防角度出发，普及冠心病基本常识及预防、治疗措施，能够有效提高群众对冠心病的认识和自我保健能力，降低患者精神负担，助力健康中国建设。

目前医学界认为影响冠心病发病的主要因素有两个方面，一是无法通过后天改变的遗传性因素，二是后天饮食、生活习惯影响的因素。从后天诱发冠心病的影响因素进行分析，主要有:(1)血压。血压的升高与冠心病发病率呈现正相关；(2)血脂异常。血脂异常时常伴随着冠心病发病率的升高，主要包括以下几方面异常:一是总胆固醇高于正常指标，二是低密度脂蛋白胆固醇高于正常指标，三是甘油三酯指标异常升高，四是高密度脂蛋白胆固醇数值过低；(3)糖尿病。糖尿病是诱发冠心病的重要危险因素之一，糖尿病患者应高度关注其冠心病的潜在发病情况，应定期检查；(4)吸烟。与不吸烟者相比，吸烟者患心肌梗死的概率增高 2～3 倍，主要有以下几方面原因:一是烟草中的尼古丁会加大心肌中的耗氧量，进而出现收缩血管和冠状动脉的情况，导致患者血压升高。二是吸烟会增加血管中的 CO 含量，从而降低血液的带氧量，导致动脉粥样硬化；(5)肥胖症。肥胖与一个人的血脂情况呈正相关关系，是诱发冠心病最重要的影响因素之一。超过标准体重 20%或体重指数(BMI)>24kg/m2 者称为肥胖症；(6)久坐不动。生活中没有运动习惯或者上班期间久坐的人，其冠心病发病率和死亡率相对于定期运动或从事外拓职业的人高 1 倍；(7)紧张。长期处于紧张的工作生活环境中，会增加冠心病的发病危险；(8)每日蔬菜或水果摄入不足。

冠心病的危险因素中，血脂异常、糖尿病等都与饮食习惯不正常、缺乏运动有关，还有很多病人生活压力过大，所以要从这些根源入手，改变一些不良的生

活方式，减小冠心病的发生概率。

二、研究目标与研究意义

1. 研究目标：
   - 探索哪些因素对十年内冠心病的发病率有较为显著的影响。
   - 对于十年内冠心病发病率的有效预测。
   - 针对建模得出的变量重要性确定预防的重点注意事项。
2. 研究意义：

研究表明，早期干预危险因素是减少心血管病负担的关键措施。全球疾病负担系列研究表明，改善上述危险因素可减少冠心病发生率 83%~89%，减少冠心病死亡率 78%~85%，因此本项目通过构建统计模型对影响冠心病发病情况的因素进行探索并得出健康指标的合理控制范围以及危险因素的影响程度大小的研究是有实际意义的。Logistic 模型可以得到影响因素具体的影响程度且能在一定置信区间下根据被调查者的各项指标对其冠心病发病情况进行有效预测；决策树与随机森林模型虽然不能得到有效的数值预测，但可以对样本数据进行很好的分类从而得出较为准确的拟合和预测结论；支持向量机模型可以在一定程度上最小化拟合误差。因此，构建上述三个模型对于有效预测冠心病发病率并提出相应预防建议是合理且合适的。

三、数据说明与变量解释

1. 研究对象与数据说明：
   - 本文的研究对象为 4238 个个体的相关医学指标水平（预测变量）与是否发生十年内冠心病（响应变量）；
   - 选取的是 kaggle 官网上 heart desease 数据集，属于医学领域数据，共包含有 15 个变量，4238 条观测。其中 TenYearCHD（十年内冠心病并发情况）为 0-1 二分类响应变量，其余为可能的影响因素，包括连续型变量和分类变量。
2. 变量解释：

对预测变量和响应变量进行数据类型展示和变量解读，如表 1：

表 1　变量说明表

变量类别	变量名称	变量类型	变量说明
因变量	TenYearCHD	分类变量	0-1 二分类变量，十年内冠心病发生情况
自变量	male	分类变量	0-1 二分类变量，0 代表女性，1 代表男性
	age	数值型变量	年龄，为连续型变量
	education	分类变量	4 分类变量，1=某些高中 2=高中或同等高中学历 3=某些大学或职业学校 4=大学
	currentSmoker	分类变量	0-1 二分类变量，0 代表目前是吸烟者，1 代表目前不吸烟
	cigsPerDay	数值型变量	连续型变量，每天吸烟根数
	BPMeds	分类变量	0-1 二分类变量，0 代表目前未服用血压药物，1 代表目前在服用血压药物
	prevalentStroke	分类变量	0-1 二分类变量，0 代表未中风，1 代表中风
	prevalentHyp	分类变量	0-1 二分类变量，0 代表不是高血压患者，1 代表是高血压患者
	diabetes	分类变量	0-1 二分类变量，0 代表不是糖尿病患者，1 代表糖尿病患者
	totChol	数值型变量	连续型变量，胆固醇水平
	sysBP	数值型变量	连续型变量，收缩压
	diaBP	数值型变量	连续型变量，舒张压
	BMI	数值型变量	连续型变量，身体质量指数
	heartRate	数值型变量	连续型变量，心率
	glucose	数值型变量	连续型变量，葡糖糖水平

数据预处理

一、缺失值模式探索

数据集共有 4238 条观测，无缺失值的完整数据共有 3656 条（如图 1）。其中 heartRate、BMI、cigsPerDay、totChol、BPMeds、education、glucose 这七个变量有缺失值。从实际意义上看，由于数据反映的是每个病人医学上的特征情况，具有较强的特异性，不太适合缺失值插补，但是从建模角度看，样本缺失值比例不低，为 13.7%，为了尽量保持数据集信息的完整性，选择 KNN 插补法，对缺失值进行处理。

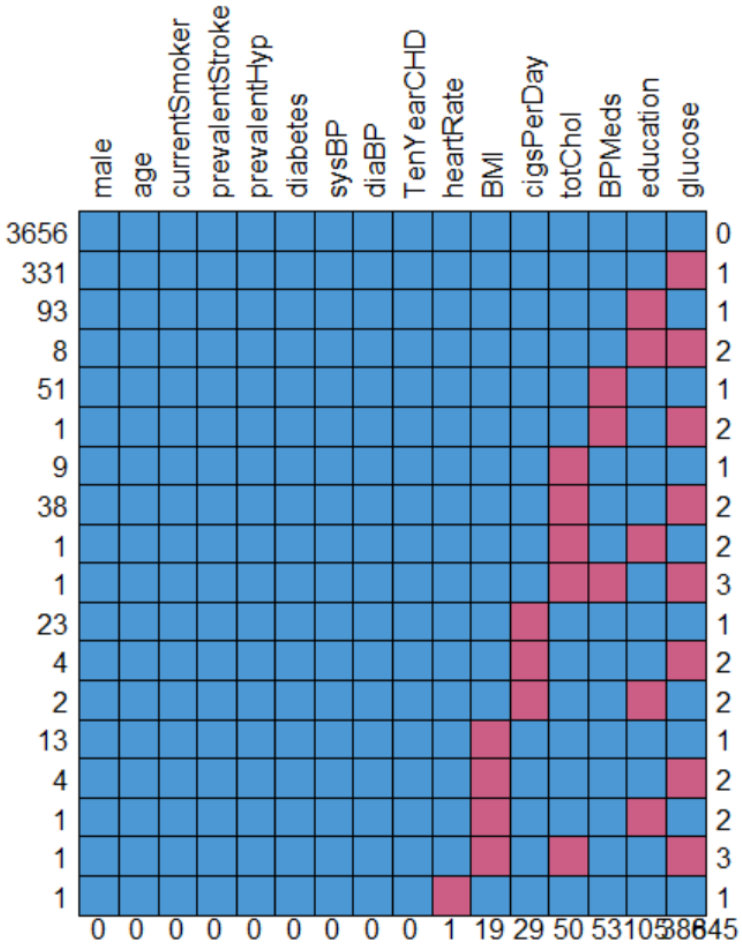

图 1 缺失值情况

二、异常值处理

对各变量画出箱线图并结合现实情况检验异常值。结合现实情况，因葡萄糖水平大于 150、BMI 大于 40、胆固醇水平大于 645 时病患基本不能存活，因此判定以上数据情况为异常值，总共 76 条观测，占比仅为 1.8%，因此剔除以上观测，最终数据集剩余 4162 条观测。具体情况如图 2、图 3、图 4 所示：

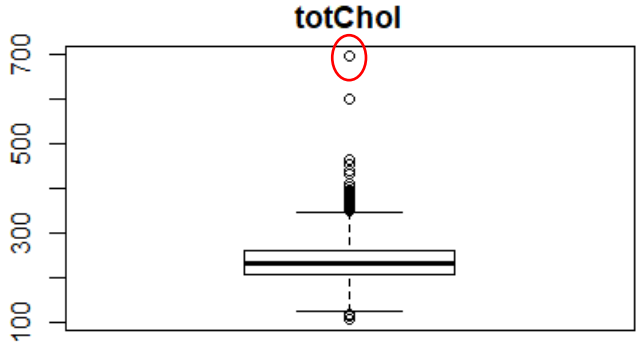

图 2　胆固醇水平异常值判定

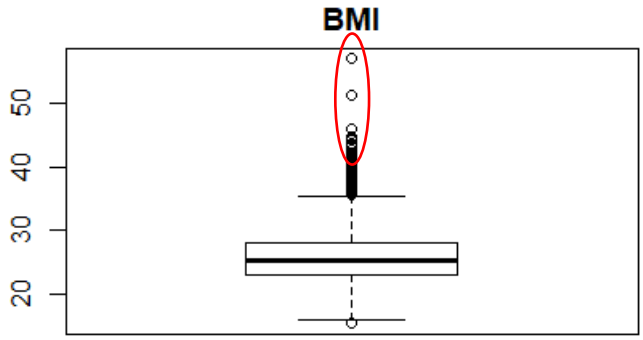

图 3　BMI 异常值判定

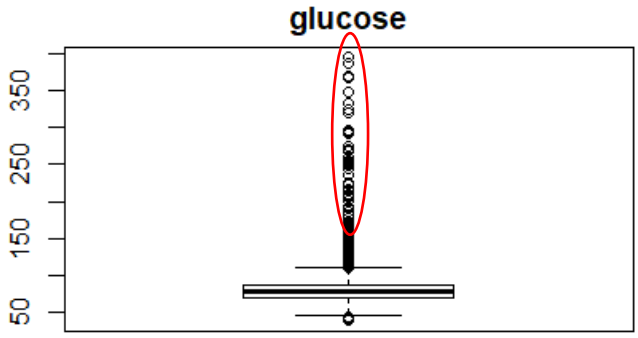

图 4　葡萄糖水平异常值判定

三、变量合并

变量 currentSmoker(当前是否为吸烟者）与 cigsPerDay（每天吸烟根数）这两列可以合并为 cigsPerDay 一列。当前不是吸烟者时，currentSmoker 和 cigsPerDay 均为 0；当前是吸烟者时，currentSmoker=1，cigsPerDay 为相应的根数。因此将两列对应行相乘，合并为 cigsPerDay 新的一列，数据与原 cigsPerDay 列的数据相同（如图 5）。

currentSmoker	cigsPerDay
0	0
0	0
1	20
1	30
1	23
0	0
0	0
1	20
0	0
1	30
0	0
0	0
1	15
0	0
1	9

→

cigsPerDay
0
0
20
30
23
0
0
20
0
30
0
0
15
0
9

图 5　变量合并

四、样本不均衡问题的处理

由于原始样本数据来自于医学领域，冠心病 10 年发病（=1）的观测数远小于未发病（=0）的观测数，响应变量 TenYearCHD=1 的比例仅为 14%。因此若不对样本数据进行均衡化处理，在后续建模时会出现分类器将异常情况也归为正常情况的可能，且会导致拟合效果和预测效果均较差。

运用 SMOTE 算法，对样本数据进行过采样和下采样，通过生成少数分类的样本、抽取多数分类的样本并使它们尽可能地均衡，从而形成新的数据集。则后续建模基于 SMOTE 算法生成的新数据集展开。

形成的新数据集 desease.ba 中共有 6743 条观测，其中响应变量取值为 1 的比例为 54.5%，样本是均衡的。这样可以减小因学习数据过少导致的分类器识别错误的概率，提高模型的准确性。

描述性统计分析

一、被调查者的基本情况

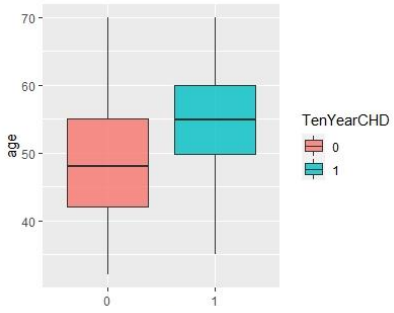

图 6 受访者年龄与是否患冠心病的分布

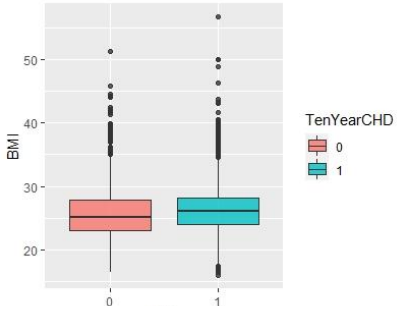

图 7 受访者 BMI 与是否患冠心病的分布

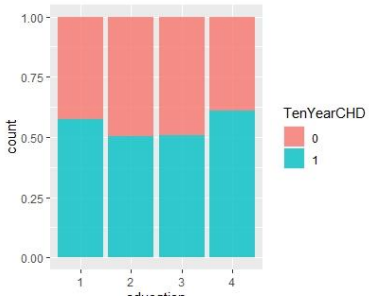

图 8 受访者受教育程度与是否患冠心病的分布

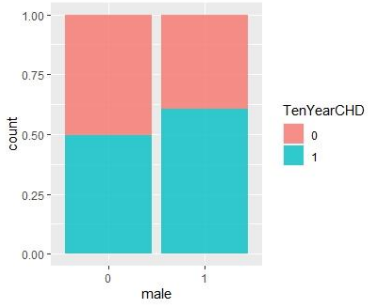

图 9 受访者性别与是否患冠心病的分布

从年龄分布来看，十年内冠心病患者的年龄总体上高于未患有冠心病的被调查者，冠心病患者年龄均值大约比未患病者高了 5-7 岁；

从 BMI 指数上看，患者和未患者的 BMI 相差不大，但总体上患者的 BMI 稍稍高于未患者；

从受教育程度来看，总体上受教育程度对患冠心病影响不大，患与未患的比例大概在 1:1，其中教育程度最低和最高的人群患冠心病的比例比中等教育程度的人群总体上稍高，高于 50%；

从性别来看，总体上男性患冠心病的比例稍高于女性，女性患与未患的比例为 1:1，男性患病比例高于 50%。

二、被调查者的健康情况

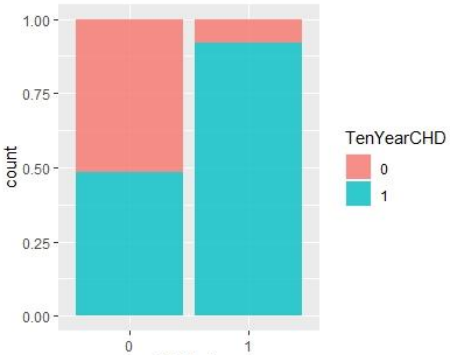

图 10 受访者是否服用高血压药物与是否患冠心病的分布

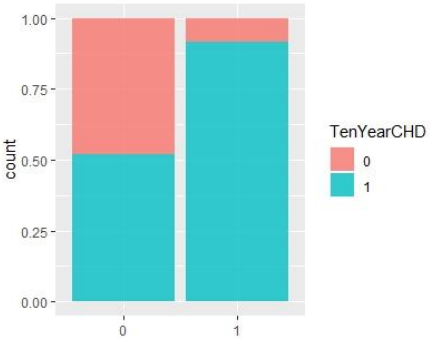

图 11 受访者是否是糖尿病与是否患冠心病的分布

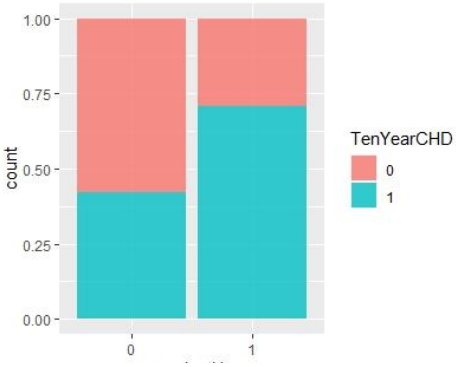

图 12 受访者是否是高血压状况与是否患冠心病的分布

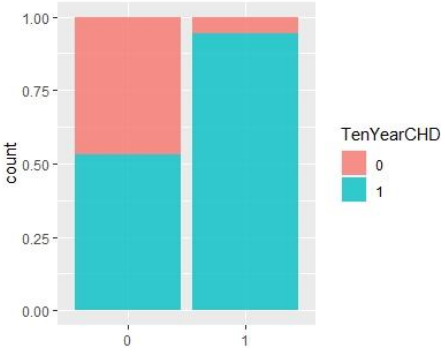

图 13 受访者是否有中风史与是否患冠心病的分布

发现患冠心病受访者中高血压状态者、服高血压药物者、患糖尿病者、曾中风患者的比例均较高。猜测这些疾病可能导致身体机能下降，从而更容易患冠心

病。

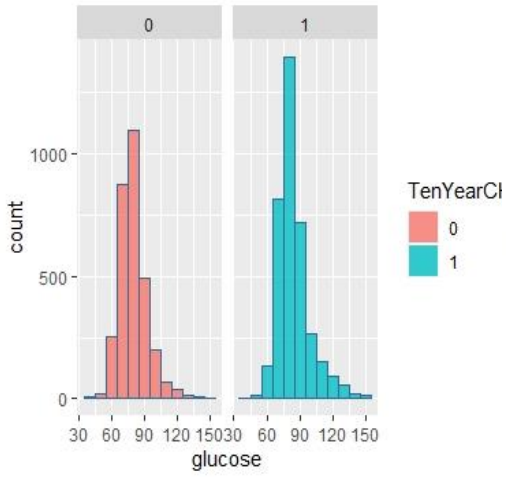

图 14 受访者血糖水平与是否患冠心病的分布

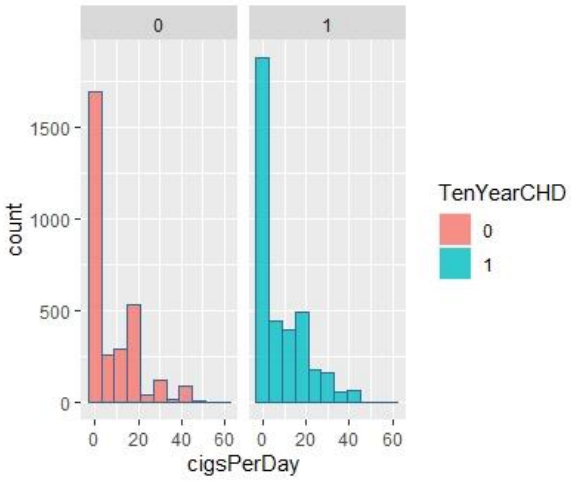

图 15 受访者每天吸烟根数与是否患冠心病的分布

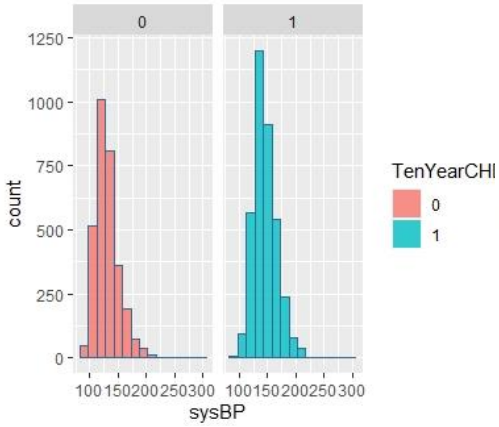

图 16 受访者收缩压与是否患冠心病的分布

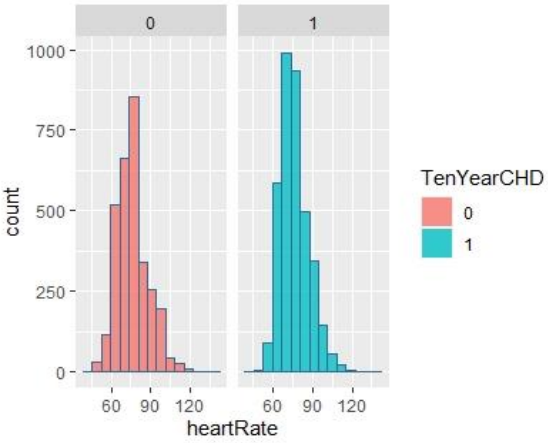

图 17 受访者心率与是否患冠心病的分布

针对分组箱线图直观差异不明显的连续型变量使用按是否患有冠心病做分面直方图。

从分面直方图 14-17 可以看出：

总体上看，十年内患上冠心病的受访者血糖（glucose）略微高于未患病受访者，二者分布形态近似一致。从心脏收缩压（sysBP）可以看出，总体上看患上冠心病的受访者心脏收缩压普遍略微高于健康受访者。而每天吸烟根数和心率则没有明显差异。

理想或正常的胆固醇水平应该低于 200mg/dl。如果读数在 200-239mg/dl 之间,就属于临界区间。如果总胆固醇水平高于 240 mg/dl,就意味是高胆固醇。

由图 18 可知，患冠心病的受访者的胆固醇水平处于正常值边界偏高的人数明显多于未患冠心病的受访者。

由图 19，患冠心病的受访者的心脏舒张压也明显高于未患者。

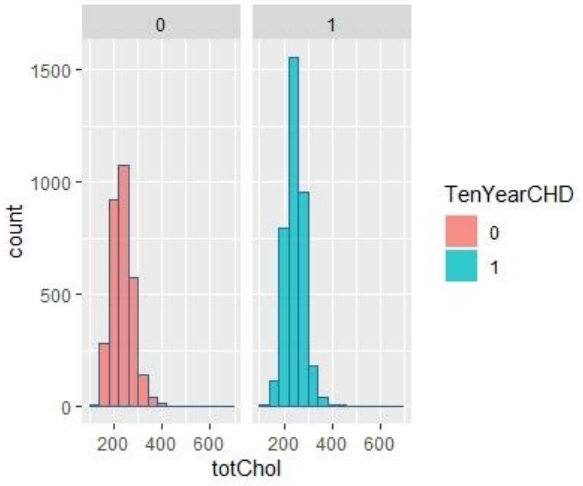

图 18 受访者胆固醇水平与是否患冠心病的分布

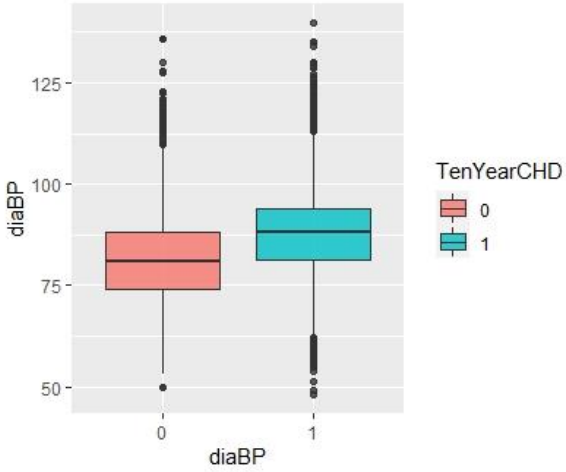

图 19 受访者舒张压与是否患冠心病的分布

模型构建

一、模型介绍及选取原因

1. Logistic 回归模型：

由于本数据集的因变量（是否患冠心病）是分类变量，如果仍采用普通的线性回归不仅误差可能不服从正态分布，而且预测值往往是连续的，预测值往往不取 0 或 1。

考虑到以上问题，对本数据采用 Logistic 回归模型：

$$\text{logit}(p(\boldsymbol{x})) = \log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

其中，模型的回归参数采用极大似然估计，对数似然函数可以写成：

通过该模型，可以根据已有自变量预测出未来 10 年是否患冠心病的概率有多大。

$$l(\beta) = \sum_{i=1}^{m} \left[y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i) \right]$$

2. 决策树与随机森林模型：

决策树是最常见的机器学习方法，当响应变量为二分类变量时，可以构建较为简单的二叉分类树模型对测试集数据进行分类学习与拟合。运用 CART 算法，利用 GINI 系数来选择划分属性。在决策树模型的构建中，较为重要的是对树模型进行前剪枝和后剪枝，通过设置 cp 复杂度、最小节点数、交叉验证次数、树的最大深度等参数构造一棵 CART 决策树；通过剪掉冗余繁杂的枝叶，使得决策树不至于分支过多同时也能展现有效分类信息。

随机森林的本质是很多棵决策树的集合，通过分类器投票产生最优模型，可以在一定程度上减小单棵决策树的误差，提高拟合优度。随机森林的模型搭建需要确定 mtry 和 ntree 两个重要参数，前者确定进入模型的最佳变量个数，后者确定森林中树的棵树。

在样本量较大的情况下，决策树与随机森林的机器学习方法对于拟合和预测效果较好，对于分类响应变量的影响因素解读更为简洁、直观、易于非统计专业的其他人士解读和理解，且决策树可视化后对于指标根据冠心病发病率影响的范围划分对控制指标在正常范围内以减小冠心病发病率具有有效预防意义。

3. 支持向量机模型：

支持向量机模型是一种二分类的模型，它的基本模型定义是使两类数据点在特征空间上间隔最大的线性分类器。如果引入非线性核函数，则可以将样本点从低维度空间映射到高维度空间，将原本非线性问题变换为高维空间里的线性问题，使其成为实质上的非线性分类器。

在分类过程中，支持向量机有一个重要的假设：每个数据的权重并不相同。除去少数几个支持向量（靠近分离超平面的数据），其他数据的权重其实等于 0。因此，支持向量机在训练时并不会考虑所有数据，而是只关心靠近分界线的数据点。因此，选用 SVM 模型就可以在该特征空间中寻找最优分类超平面。

二、Logistic 模型

首先将含有 k 个水平的分类自变量设置 k-1 个哑变量以避免虚拟变量陷阱。并且注意到自变量中有一些分类变量可能存在相关性（如是否高血压和是否服用高血压药物），但在采用列联表分析中的卡方检验之后发现 p 值<0.001，说明它们之间在 5%的显著性水平下是独立的。

之后以是否违约作为因变量，从原有数据集的观测中抽取 80%作为训练集拟合 logistic 回归，并使用逐步回归（AIC）进行变量选择，得到结果如下表所示：

表 2 logistic 回归结果 1

	Estimate	Pr(>|z|)	
(Intercept)	-7.96288	< 2e-16	***
male1	0.438192	1.49E-10	***
age	0.077612	< 2e-16	***
education2	0.066859	0.39577	
education3	0.10654	0.28407	
education4	0.542144	6.69E-07	***
cigsPerDay	0.018209	2E-09	***
BPMeds1	2.230162	< 2e-16	***
prevalentStroke1	2.286597	2.77E-13	***
prevalentHyp1	0.482407	1.79E-10	***
diabetes1	2.027786	< 2e-16	***
totChol	0.002265	0.00484	**
sysBP	0.012353	4.79E-06	***
diaBP	0.009074	0.04828	*
BMI	-0.01811	0.05517	.
glucose	0.006414	0.01221	*

可以发现 AIC 变量选择删除了 heartrate 一个变量，剩余的大多数变量都是显著的。用 deviance 检验进行检验，发现 p 值<0.001，说明在 1%的显著性水平下该模型非常显著。

注意到模型中 BMI 和 diaBP 不是很显著，因此尝试去掉这两个变量的嵌套模型，结果如下表所示：

表 3 logistic 回归结果 2

	Estimate	Pr(>|z|)	
(Intercept)	-7.96163	< 2e-16	***
male1	0.436125	1.38E-10	***
age	0.075891	< 2e-16	***
education2	0.081647	0.29773	
education3	0.124866	0.20766	
education4	0.562864	2.17E-07	***
cigsPerDay	0.018326	1.33E-09	***
BPMeds1	2.239305	< 2e-16	***
prevalentStroke1	2.315015	1.47E-13	***
prevalentHyp1	0.494664	3.86E-11	***
diabetes1	2.0119	< 2e-16	***
totChol	0.00223	0.00548	**
sysBP	0.015379	2.18E-15	***
glucose	0.006028	0.01807	*

对变量改变前后的模型进行似然比检验，得出结果 p 值=0.01159>1%，在 1%的显著性水平下不能拒绝原假设：认为原模型是冗余的。因此采用去掉 BMI 和 diaBP 之后的模型。

在该模型下：

其他条件相同的情况下，男性未来 10 年患冠心病的概率是女性的$e^{0.44}$；其他条件相同的情况下，年龄每增长一岁，未来十年患冠心病的概率是原来的$e^{0.08}$；其他条件相同的情况下，高中或高中同等学历未来十年患冠心病的概率是某些高中的$e^{0.08}$，某些大学或职业学校的受访者未来十年患冠心病的概率是某些高中的$e^{0.12}$，大学学历受访者未来十年患冠心病的概率是某些高中的$e^{0.56}$；其他条件相同的情况下，每天每多抽一根烟，未来十年患冠心病的概率是原来的$e^{0.02}$；其他

条件相同的情况下，服用高血压药物的受访者未来 10 年患冠心病的概率是不服用高血压药物受访者的$e^{2.24}$；其他条件相同的情况下，有中风史的受访者未来 10 年患冠心病的概率是没有中风史的$e^{2.35}$；其他条件相同的情况下，有高血压状况的受访者未来 10 年患冠心病的概率是没有高血压状况的受访者的$e^{0.49}$；其他条件相同的情况下，有糖尿病状况的受访者未来 10 年患冠心病的概率是没有糖尿病的受访者的$e^{2.01}$；其他条件相同的情况下，胆固醇水平每升高 1，未来 10 年患冠心病的概率是原来的$e^{0.002}$；其他条件相同的情况下，收缩压水平每升高 1，未来 10 年患冠心病的概率是原来的$e^{0.02}$；其他条件相同的情况下，葡萄糖水平每升高 1，未来 10 年患冠心病的概率是原来的$e^{0.006}$；

为了获取每个特征对未来 10 年是否患冠心病的影响，从而探究特征的重要性，采用 bootstrapping 抽样的方法，对 logistic 回归模型进行训练，通过分析每个特征下，模型的 ROC 曲线，得到下图所示特征重要性排序：

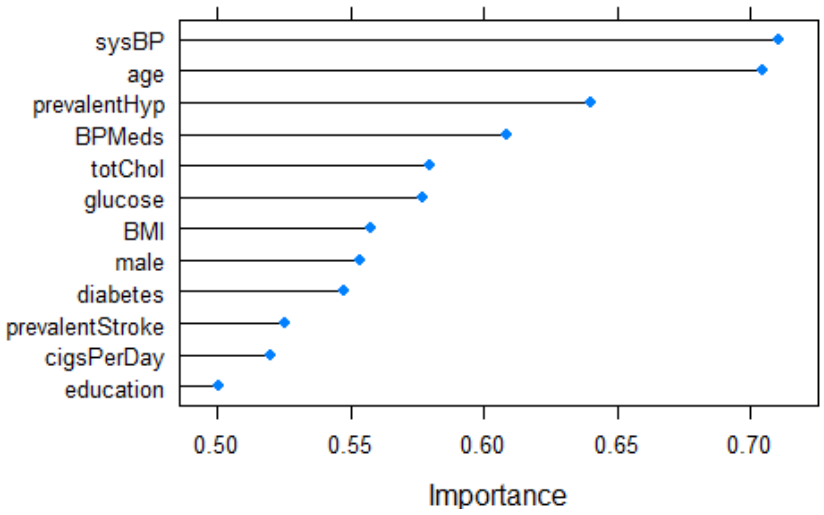

图 20 特征重要性排序图

从上图可以看出，在 logistic 回归模型中，认为 sysBP 和 age 的重要性最大，且对未来 10 年患冠心病都是正的影响，即二者的值越大，则受访者越容易在未来 10 年内患冠心病；变量 education 对未来 10 年是否患冠心病的影响十分微弱，其重要性程度接近 0.5，表明其几乎没有预测能力。

三、决策树与随机森林

（一）决策树

1. 决策树变量重要性排序：

对最大树深度、cp 复杂度以及叶子结点最大容量等调参后输出对与响应变量影响程度即重要性排序结果，如表 4、图 21 所示：

表 4　决策树变量重要性排序表

变量名	重要性
age	322.1058
BPMeds	207.7893
sysBP	135.5543
cigsPerDay	50.1736
male	47.8397
diaBP	42.2916
prevalentHyp	39.2016
totChol	26.1246
glucose	23.6927
BMI	7.1533
heartRate	2.3029
prevalentStroke	0.9028

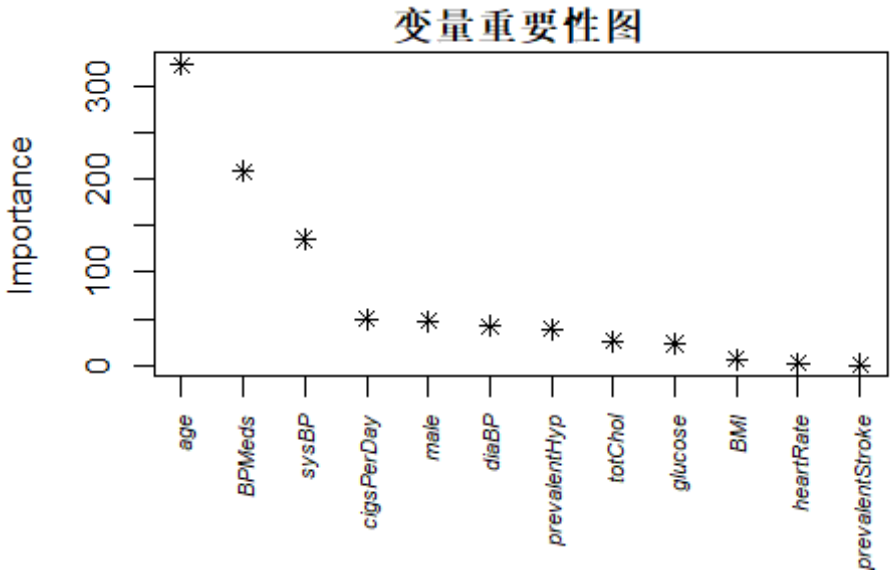

图 21　决策树变量重要性图

从表 4、图 21 可知，经过决策树的变量筛选，在决策树模型中影响冠心病发病率的较为重要的因素有年龄、是否使用血压药物、收缩压、每日吸烟根数，而是否中风对于冠心病发病率的影响相对较小。

2. 决策树模型展示与解读：

决策树的模型展示如图 22 所示：

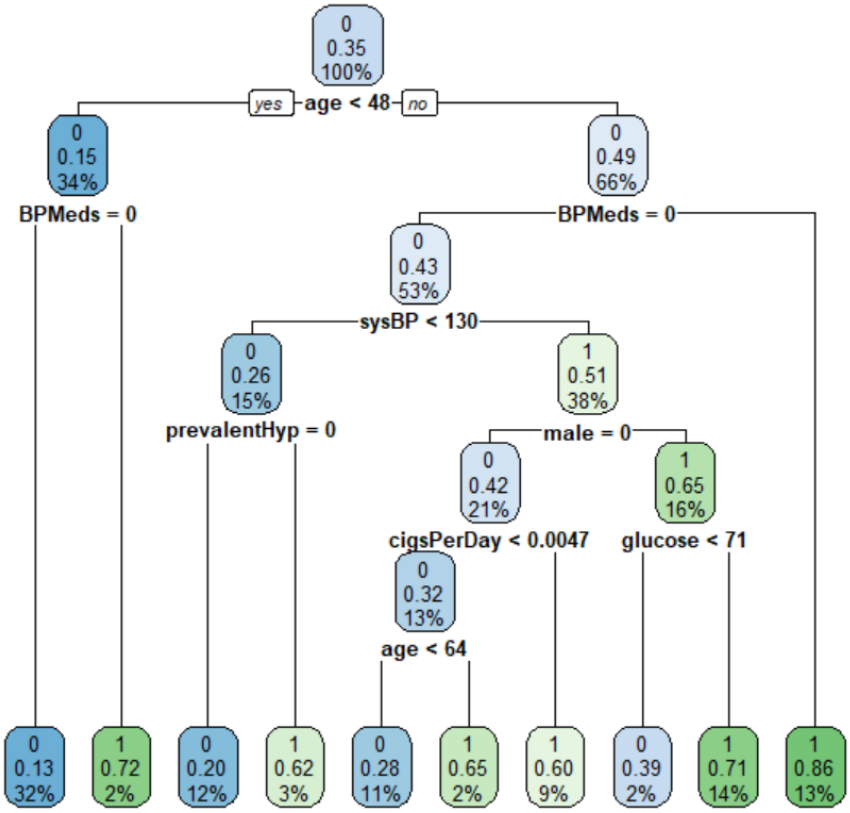

图 22　决策树可视化

从叶子节点可以得出结论，当年龄 age 大于等于 48 且目前在服用血压药物的人群是最有可能在十年内发生冠心病的，概率达到 86%，在样本中所占比例为 13%；其次是年龄小于 48 且目前在服用血压药物的人群，发生概率达到 72%，在样本中所占比例为 2%。以上两点说明，血压是导致冠心病发病的重要因素，高血压很有可能会引发冠心病的发作。

葡萄糖水平过高和高龄的人群也有较大可能发生冠心病。年龄高于 48 岁、未服用血压药物、收缩压大于等于 130、葡萄糖水平高于 71 的男性十年内发生冠心病的概率达到 71%，在样本中占比 14%；年龄小于 48 岁且未服用血压药物的人群十年内发生冠心病的概率最小，为 13%，占比 32%。

（二）随机森林

1. 随机森林模型构建：

经过模型搭建与调整，设置 mtry=5 个变量和 ntree=500 棵树，生成随机森林，调参过程如图 23、图 24 所示：

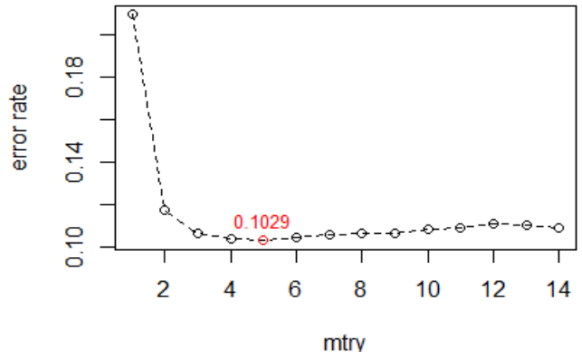

图 23　随机森林入选变量个数

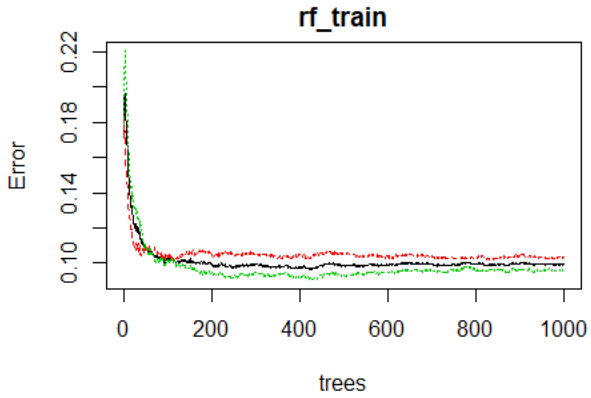

图 24　随机森林树数选择

2. 随机森林变量重要性：

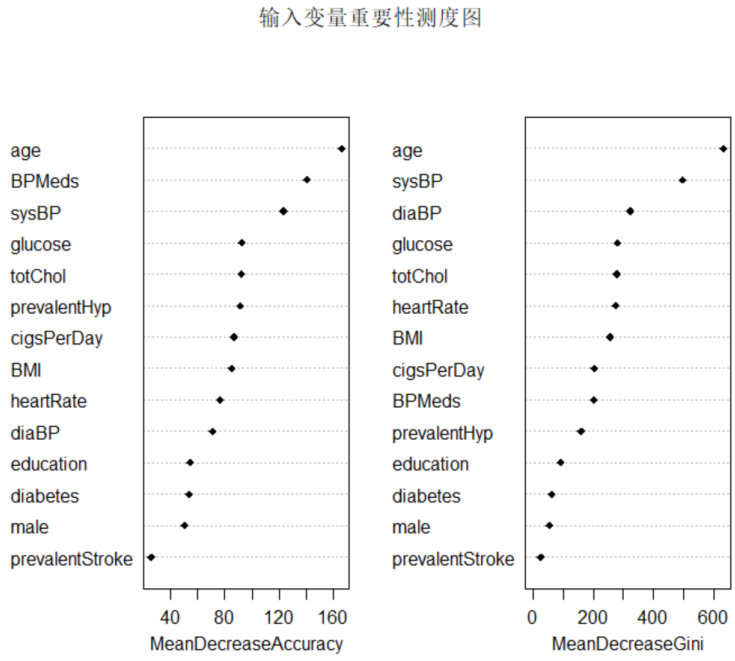

图 25　随机森林变量重要性图

从上图可以看出，无论是基于准确性下降排序还是基于基尼系数下降排序，对于冠心病发病率影响较大的变量均为年龄、收缩压、葡萄糖水平、胆固醇水平、每日吸烟根数、BMI、心率以及是否服用血压药物。综合决策树与随机森林的变量重要性情况，可以发现，年龄情况、血压指标、血糖指标、胆固醇指标以及吸烟情况对于冠心病发病的影响程度均较大，这说明，合理控制以上指标在正常范围内，对于预防冠心病具有较大作用。

四、支持向量机

1. SVM 模型建立：

首先观察数据集，样本数量远大于特征数目，因此使用非线性核函数，通过将样本映射到更高的维度，来得到更好的拟合效果。

2. 核函数选择：

在上述情况下，多项式核函数或径向基核函数较为常用。

- 多项式核函数：$\mathrm{k}(x_i, x_j) = (\gamma\langle x_i, x_j\rangle + \mathrm{c})^n$　γ: gamma; $\mathrm{c} = \mathrm{coef0}; \mathrm{n} = \mathrm{degree}$
- 径向基核函数：$\mathrm{k}(x_i, x_j) = \exp(-\gamma\|x_i, x_j\|^2)$ γ: gamma

除了需要核函数的参数之外，还要考虑惩罚系数 cost。通过控制 cost 在合理的范围内，可以表示异常分布的点对目标函数的贡献权重，cost 越大那么对异常分布点的惩罚就越大。如果选择多项式核函数，则有 gamma、coef0、degree 三个核函数参以及乘法系数四个参数要选择，调参过程比较困难；另外若多项式的阶数较高，模型复杂度会极高，导致模型求解困难、计算效率低下。因此选择径向基核函数，下面研究对惩罚系数以及 gamma 的选取。

3.参数调整（gamma、cost）：

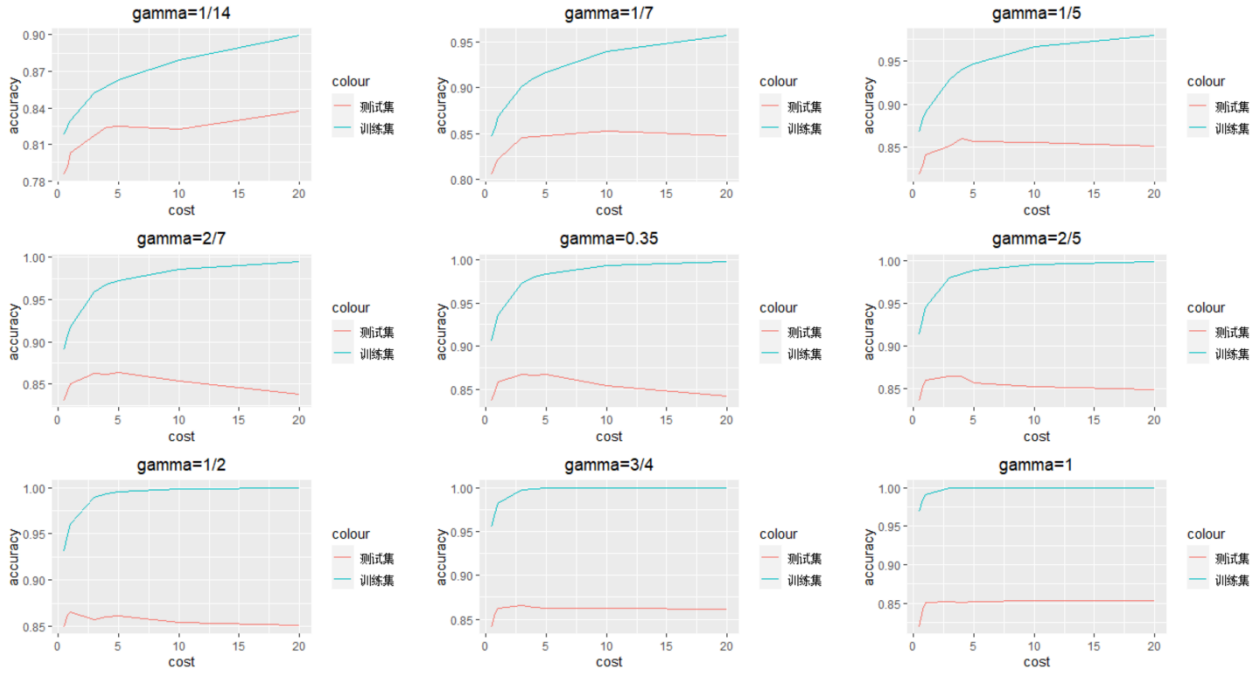

图 26　支持向量机参数调整

（1） gamma 越大对应支持向量数量越少；反之，gamma 越大，则支持向量数量越多，那么模型训练和模型预测速度越慢。

（2） 惩罚系数越高，对异常分布点的惩罚就越大，cost 过大将导致模型过拟合，反之太小会则模型欠拟合。

（3） 在构建 SVM 模型前，需要对以上两个参数进行调试，寻找最优的参数组合：对$\begin{cases} \text{cost} \in \{0.5, 0.75, 1, 3, 4, 5, 10, 20\} \\ \text{gamma} \in \{ \ \ 1/14, 1/7, 1/5, 2/7, 7/20, 2/5, 1/2, 3/4, 1\} \end{cases}$的参数组合下，以模型的错分率作为比较指标，观察发现当$\begin{cases} cost = 3 \\ gamma = 1/7 \end{cases}$时模型表现较好。其中，当 gamma 大于 1/2 时，拟合效果受到惩罚系数影响较小（如图 27）。

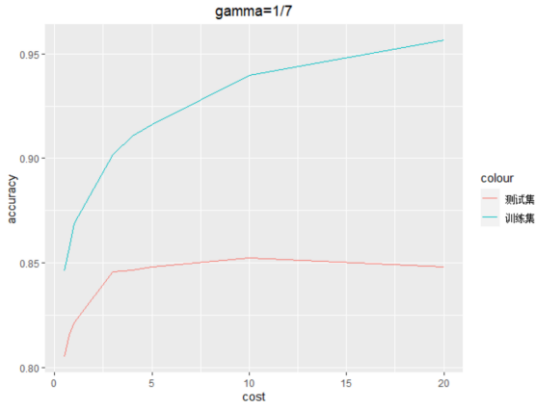

图 27 最终参数确定

4. 最终的支持向量机模型：

选择径向基核函数，在参数组合为$\begin{cases} cost = 3 \\ gamma = 1/7 \end{cases}$时，获得最终的支持向量机模型。在训练集 5394 个样本下，支持向量总数为 2649 个，因变量未来十年冠心病未发生样本 1338 个，发生样本 1311 个，分布均匀。该模型在训练集预测正确率 90.16%，测试集预测正确率 84.58%，混淆矩阵见模型评估与模型比较部分表 8-9。

模型评估与模型比较

一、Logistic 回归模型：

（一）logistic 回归模型的 ROC 曲线：

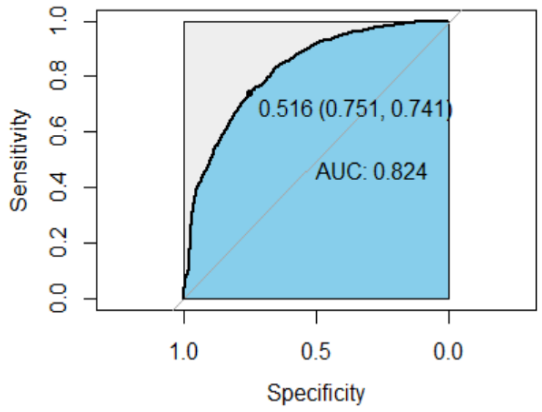

图 28 训练集 ROC 曲线

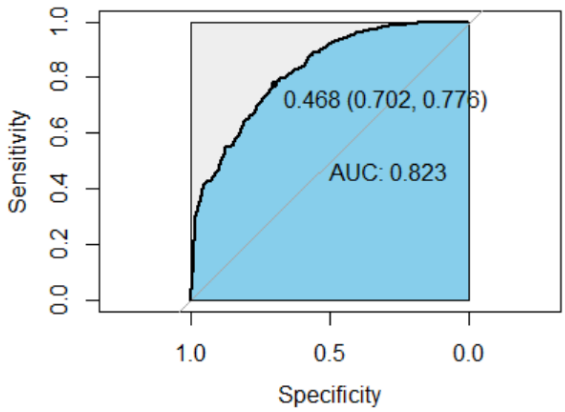

图 29 测试集 ROC 曲线

通过比较训练集和测试集的 ROC 曲线，发现二者相差不大，说明模型不存在过拟合或欠拟合的问题。并且 AUC=0.824，说明预测效果也比较好。

（二）logistic 回归模型的混淆矩阵：

表 5 Logistic 训练集混淆矩阵

训练集	0	1
0	1847	634
1	733	2180

表 6 Logistic 测试集混淆矩阵

测试集	0	1
0	422	162
1	197	568

计算得到 Logistic 模型训练集和测试集的准确率分别为 74.7%和 73.4% ，拟合和预测的准确率较高，模型较为准确。

二、决策树与随机森林模型：

（一）决策树的 ROC 曲线：

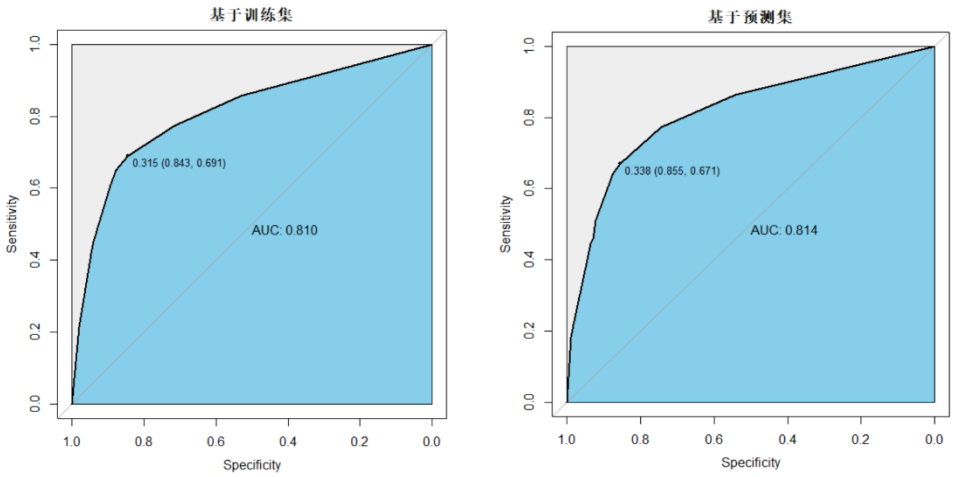

图 30　决策树训练集和测试集的 ROC 曲线

从两幅 ROC 曲线图可以看出，训练集和预测集的 AUC 值均较高，且相差不大。模型拟合效果较好且不存在过拟合和欠拟合现象。预测集的最佳阈值 P0=0.338，在此阈值下可以得到最高真阳性率和最低假阳性率。

（二）随机森林的混淆矩阵：

输出随机森林模型的混淆矩阵如表 7 所示，计算得到预测准确率为 0.9010826.模型拟合效果较好。

表 7 随机森林混淆矩阵

Pred_f	0	1
0	2749	351
1	316	3327

三、支持向量机模型：

（一）支持向量机的 ROC 曲线：

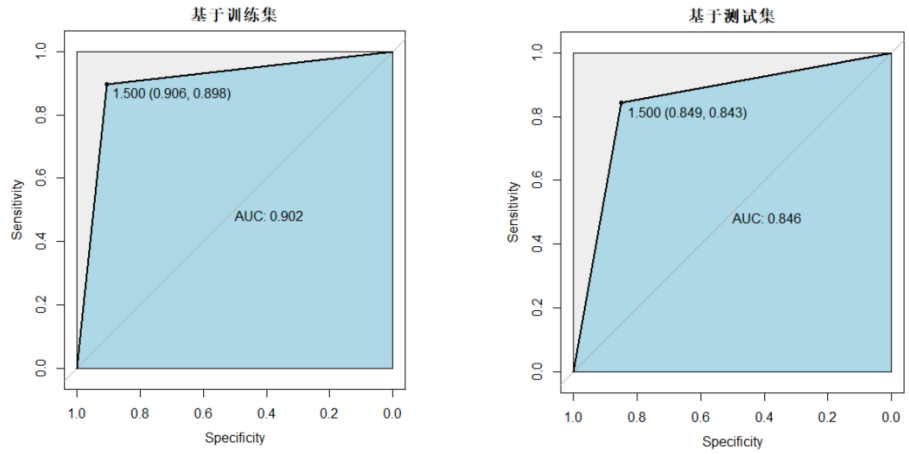

图 31　支持向量机 ROC 曲线

通过比较支持向量机在训练集和测试集的 ROC 曲线，发现二者存在差异，说明模型存在一定过拟合问题，需要引起注意。

（二）支持向量机的混淆矩阵：

输出支持向量机的混淆矩阵如表 8-9 所示，计算得到预测准确率分别为 84.58%和 90.16%模型拟合效果较好，但存在过拟合。

表 8 SVM 训练集混淆矩阵

训练集	0	1
0	2248	298
1	233	2615

表 9 SVM 测试集混淆矩阵

测试集	0	1
0	496	120
1	88	645

SVM 模型优缺点总结：支持向量机模型通过径向基核函数向高维空间进行映射，在使样本与分界超平面的间隔最大化过程中，支持向量对确定分界面起决定作用，而非样本空间的维数,某种意义上避免了“维数灾难”。但在大样本数据上运算速度较慢，且对参数与核函数选择敏感，虽然经过参数调整，本模型仍存在一定的过拟合。

四、模型诊断与比较

为了选择最优的模型，以达到预测未来十年冠心病发病与否，针对测试集准确率、训练集准确率、AUC 对三种模型进行比较，得到汇总结果表如下：

表 10 三种模型比较

	逻辑回归	随机森林	支持向量机
测试集准确率	74.65%	88.88%	84.58%
训练集准确率	73.38%	90.11%	90.16%
AUC	0.824	0.814	0.843

比较模型在测试集和训练集上的表现，虽然逻辑回归模型泛化能力最强，但是预测准确率最低。而支持向量机模型虽然提高了模型的预测精度、有最高的 AUC 但模型泛化能力明显削弱，随机森林模型兼具了两者的优势，预测精度较高、泛化能力较强，但 AUC 较低。上述比较没有得到明显最优的模型，因此做进一步比较。

五、学习曲线与模型选择

通过五折交叉验证，得到三种模型 AUC 学习曲线，总体来看随机森林表现最好。

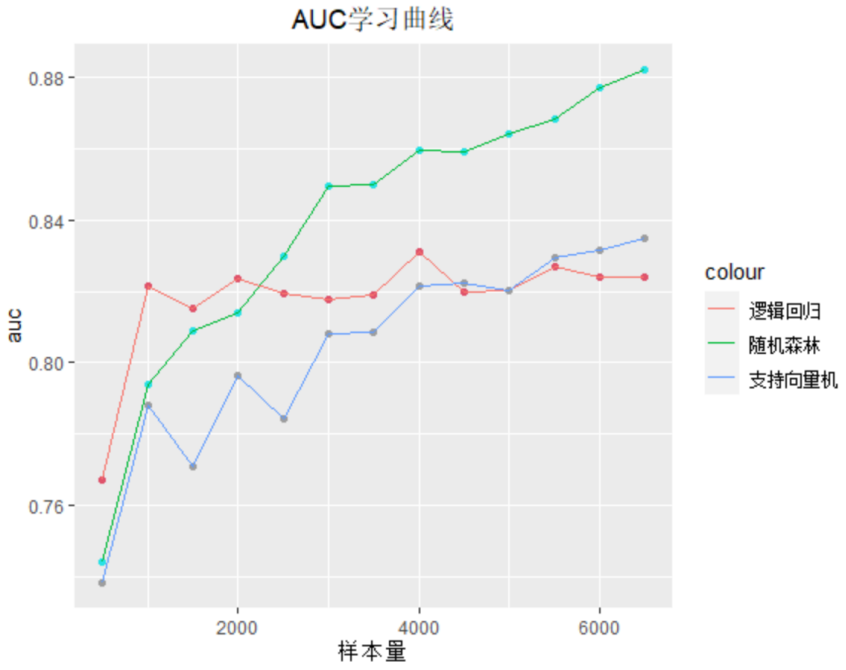

图 32 三种模型的学习曲线

通过图 32 的学习效率比较，可以得出以下结论：

- 随机森林 AUC 与样本数量成正比，且还有继续增加趋势；
- 逻辑回归在小样本情况下效果较好，样本数超过 1000 后 AUC 稳定在 0.82 左右；

- 支持向量机则介于两者之间

结合之前对每个模型本身的优劣势分析，由于医学数据样本相对珍贵，难以采集。当样本量有限（小于 1000 时），可以采用逻辑回归模型，相对预测能力较好。当样本量足够大时，我们首先引入了决策树模型，决策树模型有简单，逻辑清晰，可解释性强的优点，但容易过拟合。于是我们引入随机性，采用随机森林，用多个决策树的投票机制，有效弥补了决策树容易过拟合的缺陷；虽然其优点在于简单、容易实现、计算开销小，而且在实际任务中表现良好，但不足的是随机森林内决策树之间其实是独立的，也就是说每棵决策树都是一颗孤立的树，没有对周围的树产生正面影响。因此尝试支持向量机模型，但经过参数调整，仍然存在过拟合问题。综上，在大样本下，我们倾向于选择随机森林模型来预测。但为给受访者提出合理建议，采用 Logistic 模型和决策树模型可以得到应当特别关注的健康指标及其正常范围，更有实际意义。

结论

从 logistic 模型来看，高血压、高血糖、高胆固醇的身体状况极容易提高未来 10 年冠心病的患病率；从决策树和随机森林模型来看，收缩压、葡萄糖水平、胆固醇水平、每日吸烟根数、BMI、心率以及是否服用血压药物这些健康指标对于冠心病患病情况影响较大。

此外，可以通过 logistic 模型预测人们未来 10 年患冠心病的概率，从而划分出患病概率较高的人群，对该类人群进行重点监控防护，并采取有效预防措施。综上所述，本报告建议受访者在保证自身其他健康指标在正常范围内的情况下，特别关注自身的血压、血糖、胆固醇水平、葡萄糖水平以及 BMI，平时饮食尽量清淡、避开多油多盐的食物，少吃甜食，多吃粗粮和蔬菜水果等，少饮酒少吸烟，养成规律健康的运动习惯。

参考文献

[1]陈斌.临床实践视角下的冠心病预防和治疗[J].海峡科学,2020(10):64-67+72.
[2].预防冠心病，年轻人不可忽视[J].江苏卫生保健,2020(11):19.

附录：小组分工

吕郢歌：课题综述、样本不均衡处理、决策树与随机森林建模；

李晨茜：数据预处理、描述性统计分析、Logistic 建模；

范怡雯：SVM 建模、三种模型比较与选择。

PPT、pre、论文均大致按照上述分工进行，论文除以上个人部分，其余部分合作完成。