

基于数据挖掘的心脏病风险预测与分析

目录

- 任务介绍
- 课设要求
- 时间场地
- 考核形式
- 报告要求

任务介绍

心脏病是指影响心脏和血管系统的各种疾病，最常见的心脏病类型包括冠心病（即心脏供血不足导致的心肌损伤）、心绞痛等。心脏病通常与生活方式、饮食习惯、遗传因素、环境影响等多种因素密切相关。

根据世界卫生组织的统计，心脏病是全球致死的主要原因之一，特别是在发达国家和一些发展中国家，心脏病导致的死亡率呈上升趋势。此外，患有心脏病的人往往面临长期的病痛折磨，可能无法从事高强度的体力活动，甚至影响日常生活和工作能力，显著影响生活水平。

而心脏病的发生与许多可调节的危险因素相关，如高血压、高胆固醇、吸烟、肥胖、缺乏运动等。因此，提前识别出高风险群体并进行干预，可以显著降低心脏病的发生率和死亡率。

本次课程设计旨在探索利用数据挖掘技术和统计分析方法，预测个体心脏病的风险，分析影响心脏病发生的主要因素，并为更好地管理和控制心脏病提供科学依据。

任务介绍

提供的数据集中包含了与心脏病预测相关的多个属性。这些属性涉及个体的健康状况、生活方式等方面，为构建心脏病风险预测模型提供了丰富的原始数据。

每个属性的含义可以参考 [Codebook2022.HTML](#)。

属性				
State	Sex	GeneralHealth	PhysicalHealthDays	MentalHealthDays
LastCheckupTime	PhysicalActivities	SleepHours	RemovedTeeth	HadHeartAttack
HadAngina	HadStroke	HadAsthma	HadSkinCancer	HadCOPD
HadDepressiveDisorder	HadKidneyDisease	HadArthritis	HadDiabetes	DeafOrHardOfHearing
BlindOrVisionDifficulty	DifficultyConcentrating	DifficultyWalking	DifficultyDressingBathing	DifficultyErrands
SmokerStatus	ECigaretteUsage	ChestScan	RaceEthnicityCategory	AgeCategory
HeightInMeters	WeightInKilograms	BMI	AlcoholDrinkers	HIVTesting
FluVaxLast12	PneumoVaxEver	TetanusLast10Tdap	HighRiskLastYear	CovidPos

课设要求

本次课设主要完成以下两个任务：

■ 心脏病风险预测

利用其他属性**预测**“HadHeartAttack”，从而判断该个体是否患有心脏病。需要先对数据进行探索分析，包括数据可视化、缺失值处理、数据集划分、数据归一化、分类变量编码等。然后选择合适的模型进行训练并调优，最后评估模型并进行改进。

■ 心脏病相关的预防措施分析

利用该数据集深入挖掘影响心脏健康的关键因素，给出**可靠、可理解的防护措施**，比如探索心脏病和生活方式的关系、发现不健康的行为模式、分析健康状况与心脏病的关联等，发现帮助公众更好地管理自己的健康。

课设要求

任务要求：

■ 心脏病风险预测

- 构建合适的特征；
- 模型要优中选优，包括对比实验和消融实验；
- 考虑类别不平衡的问题。

■ 心脏病相关的预防措施分析

- 自选角度，至少从两个方面给出防护措施的建议；
- 建议应该具有较高的可信度。

时间场地

□ 课程设计的场地安排

化成楼124为本次课程的专用教室，推荐各位同学在这里完成课程，方便组内讨论交流。但是本次课设不强求场地，**团队可根据情况自选**。

□ 课程设计的时间安排

- 2025年2月17日之前，自行组队，每组3-4人，自拟队名；
- 2025年2月17日上午8点，在化成楼124进行课设介绍；
- 2025年2月19日上午8点，在化成楼124，每个团队进行进度汇报；
- 2025年2月21日上午8点，在化成楼124，团队正式汇报答辩（自带笔记本）；
- 2025年2月28日，课程设计提交材料（实验报告、进度汇报PPT、代码）截止。

考核形式

考核项目	分数占比	详细说明
过程性考核	5%	进度汇报，准备PPT。
课设报告	40%	任务一，主要考察数据挖掘过程的完整性。
	15%	任务二，主要考察所提供建议的可行性和有效性。
	10%	课设报告规范性，按要求书写（可自行增加模板小节），每个章节都会影响最终分数。
代码	10%	答辩时可能要按需求现场运行，随课设报告提交。
最终汇报	20%	可单人汇报或多人汇报，每组 最长20分钟 。准备好代码， 可能随时会要求解释代码实现 。
总计	100%	以团队的整体表现和队员的协作情况给出最终分数， 每个队员应各尽其能 。

报告要求
