1 Problem Background and Related Research Overview

大背景：健康+CVD+

*现代社会最致命的敌人是心血管疾病。（引用）作为一种**涉及心脏和血管系统的严重健康问题，CVD以其高发性和严重性，在全球范围内造成了大量的死亡和残疾。~~据世界卫生组织（WHO）的数据，CVD是导致全球死亡的主要原因之一，占据了世界死亡人数的显著比例。~~*

*CVD的发病率与现代生活方式的变化、工作压力的增加、环境污染以及不良生活习惯等因素密切相关。在现代社会中，快节奏的生活和不健康的生活方式使得更多人面临心血管疾病的风险，因此，正确、迅速的诊断至关重要，much care should be taken to 引入机器学习方法, 探求数据中隐藏的模式和规律，辅助医生正确诊断，**降低误解和误诊率，为CVD患者提供更优质的医疗服务。*

[1]

The modern world has cardiovascular disease as its deadliest enemy. (Prediction of Cardiovascular Disease Using Machine Learning Algorithms)As a severe health issue involving the heart and vascular system, cardiovascular disease (CVD), has resulted in a substantial number of deaths and disabilities globally, characterized by its high incidence and severity.

The incidence of cardiovascular disease (CVD) is closely associated with changes in modern lifestyles, increased work-related stress, environmental pollution, and unhealthy habits. In contemporary society, the fast-paced and unhealthy way of life exposes a greater number of individuals to the risk of cardiovascular diseases. Therefore, accurate and prompt diagnosis is crucial. Much care should be taken to introduce machine learning methods, aiming to uncover hidden patterns and regularities within data. These methods assist healthcare professionals in making accurate diagnoses, reducing misunderstandings and misdiagnoses, ultimately providing higher quality medical services for individuals suffering from CVD.

~~目前，生物入侵是农业、粮食生产和生物多样性领域关注的全球问题。入侵物种对新殖民地区有不利影响，如当地物种群落的贫困，以及提供食草动物控制和授粉服务[1]的关键昆虫的减少。~~

~~最近在加拿大和美国发现的亚洲大黄蜂（亚洲大黄蜂）引起了人们对它可能成为入侵物种的担忧。在美国昆虫学会决定柑橘属植物的官方通用名称之前，我们被建议使用“麻雀大黄蜂”[2]。小黄蜂倾向于在低山丘陵、低地森林或城市景观[3]。麻雀大黄蜂是美国的一种检疫害虫，因为它会给养蜂人[4]造成重大损失。因此，应非常谨慎地预测大黄蜂随时间的迁移，并确定人工报告，以更有效地部署政府资源来调查这个问题。~~

~~Nowadays, biological invasions are a global concern in agriculture, food production and~~

~~biodiversity. Invasive species have detrimental effects on the newly colonized areas, such as~~

~~impoverishment of local species assemblages, as well as the decline of critical insects providing~~

~~herbivore control and pollination services [1].~~

~~The recent finding of Vespa mandarinia (Asian giant hornet) in Canada and the USA has~~

~~prompted concern that it could become an invasive species. Until the Entomological Society~~

~~of America decides on the official common names for V. mandarinia, we are suggested to use~~

~~‘sparrow hornet’ [2]. Sparrow hornet tends to nest in low mountain foothills, lowland forests~~

~~or green space in urban landscapes [3]. Sparrow hornet is a quarantine pest for the USA for it~~

~~will cause a significant loss to beekeepers [4]. Therefore, much care should be taken to predict~~

~~the migration of hornets over time and identify manual reports to deploy government resources~~

~~more efficiently to investigate the problem.~~

1.1 Problem Background

~~The team’s solution should be articulate, concise, and organized in order to allow the~~

~~reader to easily follow the solution process and conclusions. Key statements should~~

~~present major ideas and results.~~

~~• A Table of Contents assists the reader in previewing the organization of your~~

~~report.~~

~~• Present a clarification or restatement of the problem as appropriate.~~

~~• Present a clear exposition of all variables and hypotheses.~~

~~• State and justify reasonable assumptions that bear on the problem.~~

~~• Present an analysis of the problem, motivating or justifying the model being~~

~~used.~~

~~• Summarize derivations, computations, or illustrative examples in the main body~~

~~of the solution, and leave lengthy derivations and/or calculations and data in~~

~~appropriate appendices.~~

~~• Include a design of the model. Discuss how the model could be tested, to include~~

~~error analysis, sensitivity, and/or stability.~~

~~• Discuss any apparent strengths or weaknesses to your model or approach.~~

~~• Provide a conclusion and report results explicitly.~~

~~• Document resources and references.~~

1.2 Related Work

For mathematical modeling beginners, the book named ‘A First Course in Mathemat-

ical Modeling’ [1] is a good reference.

【2】

在心血管疾病（CVD）研究领域，以探讨该健康问题的病因、诊断和治疗为主题的许多研究已经为我们提供了清晰的方向引领：

\item

+ CVD流行病学：

先前研究广泛调查了CVD的全球患病率==清晰指标(判断的指标)

+ CVD疾病预测中的机器学习应用：

在CVD的临床决策辅助开发领域，研究主要集中在机器学习模型与信息系统集成的混合模型领域，并不断开发、选择、优化模型，以确保在真实医疗环境中的可行性和稳健性，实现高精准度和高效率的目标。

通过多种机器算法对心血管疾病的预测研究显示出不同程度的准确性，优化研究正在不断进行中。

主要研究集中在应用模型集成方法，改进集成技术，如【1】采用包括朴素贝叶斯、随机森林（RF）等多种分类器的集成模型，使用了装袋和提升的集成技术，成功改善弱分类器的准确性，再利用多数投票方法，通过整合不同机器学习算法的结果，将心血管疾病预测模型的准确性提供到了~~85.48%。~~

此外，一些研究侧重于精准识别心血管疾病的关键风险因素，应用包括K-NN、SVM等机器学习模型，综合比较分析，混合模型连同所选属性实现了~~87.41%~~的准确性。

另一些研究则通过改进特征选择算法，结合分类模型技术，选定特征子集，提高模型性能，MFSFSA和SVM的组合成功将预测模型提升到了~~81.19%~~的准确性。

逻辑回归、knn、决策树、随机森林、svm

（文献综述）

+ 机遇与挑战 （创新性）

**Our work && model overview**

**画图--**

【3】

2 Assumptions and Data Preprocessing

2.1 Assumptions

2.2 Data Preprocessing

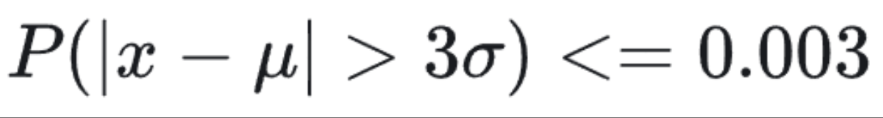
前言：

+ 数据转换：采用最小-最大归一化方法，运用（）将数据在同一量纲下线性缩放到[0,1]区间，消除量级差异，提高鲁棒性，平衡特征尺度，增强可视化效果，如图所示。

+ 数据清理：基于附件中的files cardio\_train.csv，对数据中包含的噪音进行筛查和清洗，解决数据的不一致性，降低由于传感器故障、人工录入错误或异常事件对模型检测产生的影响。

++噪声数据处理：我们首先人工剔除了不符合逻辑的数据数据，减少对模型质量的影响，尽管这些数据的占比不大。

++离群值处理: 首先通过箱线图进行可视化统计分析，快速定位异常值。鉴于异常点的数量占总数据的边缘份额，影响边际，且在统计学上不具备显著性，故可直接删除离群记录；

**结合3σ**原则，定义（）范围内的点为离群点，检查、完善离群点的清理。（引）

