上海大学

SHANGHAIUNIVERSITY

毕业论文（设计）

UNDERGRADUATE THESIS (PROJECT)

题目：基于机器学习的子宫肉瘤患者的临床病理及预后分析

学院	计算机工程与科学学院
专业	计算机科学与技术
学号	19120171
学生姓名	黄奕恺
指导教师	沈文枫
起讫日期	2023.02.21– 2023.06.03

目 录

基于机器学习的子宫肉瘤患者的临床病理及预后分析

摘 要

子宫肉瘤是源于子宫体部位的一组独立的高度恶性肿瘤，根据原发部位又分为子宫平滑肌肉瘤、子宫内膜间质肉瘤等。该肿瘤属于较为罕见的肿瘤，通常发病于45岁以上的绝经后妇女，虽然已经有一些相关文献对其进行了科普与介绍，关于其临床病理及预后分析的研究依然较少。本文基于SEER癌症数据集使用KM曲线与Cox多因素比例风险回归模型对影响其患者生存率的各个因素进行了分析与总结；并根据其预后相关数据进行了多种聚类算法的使用，分析聚类后得到簇所代表的表型具有的特征，从而分析得到子宫肉瘤患者的不同表型特点。由于聚类算法属于无监督学习，难以运用在新的数据上。所以本文建立了相关预后模型，使用决策树与随机森林对SEER上的聚类得到的表型进行训练，从而得到能区分患者所对应表型的预后模型，从而更好地评估患者的生存情况。最后结合医院提供的数据对上述模型进行了验证，总结全文工作与创新点，并展望后续工作。

关键词：机器学习, 比例风险回归模型, 决策树, 随机森林, 层次聚类

Machine learning-based clinicopathological and prognostic analysis of patients with uterine sarcoma

ABSTRACT

Uterine sarcoma is an independent group of highly malignant tumors originating from the body of the uterus, which are classified into smooth muscle sarcoma and endometrial mesenchymal sarcoma according to the site of origin. It is a rare tumor that usually develops in postmenopausal women over 45 years of age. Although it has been popularized and introduced in the literature, there are still few studies on its clinicopathological and prognostic analysis. In this paper, we analyzed and summarized the factors affecting the survival rate of patients based on the SEER cancer dataset using the KM curve and Cox multifactor proportional risk regression model; and used various clustering algorithms to analyze the characteristics of the phenotypes represented by clusters after clustering to analyze the different phenotypic characteristics of patients with uterine sarcoma based on their prognosis-related data. Since clustering algorithms are unsupervised learning, they are difficult to apply on new data. Therefore, in this paper, a relevant prognostic model is developed and the phenotypes obtained from clustering on SEER are trained using decision trees and random forests to obtain a prognostic model that can distinguish the phenotypes corresponding to the patients and thus better assess the survival of the patients. Finally, the above models were validated with the data provided by the hospital, summarizing the full work and innovations and looking forward to the follow-up work.

Keywords: Machine learning, proportional risk regression models, decision trees, random forests, hierarchical clustering

第1章　绪论

本章主要介绍了子宫肉瘤研究的相关研究的研究背景及意义，分析了相关课题的研究方法与现状，最后列举了本文的研究目标内容与结构。

§1.1　研究背景与意义

作为一种较为罕见的肿瘤，子宫肉瘤相关死亡率在子宫恶性肿瘤中中的比例在16%以上[1]，目前的主要治疗方式是以手术为主，手术后一般会通过化疗（少数情况使用放疗）辅助，以确保最后效果。根据SEER数据分析显示，子宫肉瘤患者的中位年龄在65岁左右，发病原因有很多，根据对文献中的数据显示，子宫肉瘤可能与某些遗传基因的突变有关。而子宫肉瘤的复发情况较为常见。子宫肉瘤的复发是指在手术和治疗完成后术后阶段肿瘤重新生长并再次出现的一系列过程。子宫肉瘤的复发并不局限于原发病灶，也可能出现在周边组织内，包括子宫、输卵管、卵巢等等。按照文献显示，I II期患者5年生存率为59%，而III期则为22%，IV期为9%。[1]从数据中可以看出，子宫肉瘤预后结果中，复发的概率较高，预后的效果较差。在预防子宫肉瘤的复发中，一些研究表明了术后辅助治疗和放疗能够达到降低复发率的作用，并提高患者的生存率。然而作为一种较少见的恶性肿瘤，目前子宫肉瘤方面尚未有比较公认的预后处理模型出现。为了处理近年各个医院积累的子宫肉瘤的随访与预后数据，利用数据对患者术后复发状况或生存率进行预测与评估，能够更好地帮助患者了解自身身体状况，也能帮助医生对于危险程度较高的患者进行重点关注，从而能够让医疗资源能够更加有效地被利用。因此，本文旨在建立一套使用患者预后数据训练而成的预后模型，并建立相应的UI界面，让医学人士能够方便地管理并使用该系统对于患者的将来的生存状况进行评估。

§1.2　子宫肉瘤研究发展现状

目前，子宫肉瘤的研究尚且不是很多，有很多种ML模型可以在其中子宫肉瘤的检测分析中起到一定作用，比较常用的模型主要分为以下两类：图像模型、数据模型。

§1.2.1　在子宫肉瘤方面使用的图像模型

图像模型主要用于子宫肉瘤的诊断与治疗方案的提出。对于使用图像进行的分析中，CT和MRI是主要使用的方法。图像模型主要通过人工智能技术对于图像进行

分隔，提取出ROI，并用于分类模型。分类模型是用于分类和预测的模型，目前常用于子宫肉瘤的治疗决策中。其中，Transformer与生成对抗网络（GAN）是目前比较常用的模型。Transformer可以实现多模态的医学图像分类。以往使用的深度神经网络由于是基于卷积架构形成，它在图像像素较为清晰，或内容较为复杂时，难以对于图像中结构的远端依赖性有较为明确的认知，对于复杂情况分类效果并不好。然而在使用了自注意力机制的Transformer后，它对远端结构的编码让它拥有了更强的学习表达能力，从而有了更好的分类效果。而生成对抗网络则可以生成与真实数据相似的模拟图像，通过模拟图像与真实图像的对比，生成模型与判别模型的不断迭代提升。GAN生成的图像可以为有限的图像数据添加标注后的新数据，同时其附带的判别模型也可以用于对于医生训练数据的扩充。

§1.2.2 在子宫肉瘤方面使用的数据模型

最常用的分类模型中，支持向量机和随机森林首当其冲，这些模型可以用于预测子宫肉瘤的侵袭能力和转移风险，从而为临床医生提供重要的治疗决策参考。

§1.3 本文的研究内容及目标

§1.3.1 研究内容

本文旨在设计并使用Vue与Flask作为前后端技术栈构建一个简易的患者病情预测平台，调用Python实现的预后模型。从而可以帮助医生预判患者病情，使得医生倾注医疗资源来为高风险患者进行进一步的病情随访，推动医学诊断的数字化，让医生能在这个更便携的平台上开展一系列工作，同时也为医疗服务的集成提供了一条较为有效的工程实践经验。

§1.3.2 研究目标

对于本文的研究内容，我制定了以下几条目标：

1） 从SEER数据库获得数据，验证各类模型的前置条件是否满足，测试各类模型方法，评估各类方法的作用、优缺点与效果。

2） 实现一个具有登录、授权功能的前后端系统，添加工单功能，同时引入数据导出与导出功能，在后端中集成Python实现的预后模型方法，并提供在线的训练与预测功能，从而能够满足医生与管理员的共同使用。

§1.4 本文组织结构

整篇论文一共分为七章。

第一张介绍了子宫肉瘤相关研究的背景与意义，阐述了当前子宫肉瘤相关研究的内容与方向，并说明了本文的研究目标与具体内容。

第二章主要介绍了本文使用的代码环境与技术，描述了相关的曲线或参数的具体含义，并阐释了为何使用这些技术与指标。

第三章是本文筛选并处理数据集的过程，描述了本文中如何从SEER癌症数据集与医院数据集中分别下载并处理数据，从而能在下面的章节中分析并使用。

第四章分析了数据集中的内容，利用数据进行了生存分析，使用KM曲线与Cox比例风险模型研究各影响因素对生存的影响，并进而验证数据的有效性。

第五章首先使用无监督聚类模型对患者的表型进行分类，并分析了各表型所具有的不同特征与预后效果的不同。用分类得到的患者表型数据作为监督学习的数据源，使用决策树与随机森林模型以判断患者所处的表型，并与使用三年生存率的传统方法进行了对比，评估了两者的效果。

第六章介绍了根据模型实现的系统主要功能及实现过程中的技术细节。

第七章对全文进行了总结，归纳了本文的创新点与具体内容，并指出了本文使用的模型的局限性与改进方向。

第2章　子宫肉瘤预后模型相关技术综述

本文的子宫肉瘤预后模型使用Python搭建模型，力求模型能够具有更好的可移植性，并依据这一点构建了相关系统。系统主要使用Vue.js作为核心框架，并配合以TypeScript和Sass作为技术补充；服务端考虑到模型算法基于Python，使用轻量级并高度可定制的Flask实现，并使用Docker容器技术让后端的部署更加稳定、便携。

同时，本文亦使用了一些医学方面的常用模型与分析方法，本章主要对使用到的相关技术、生存分析方法与预后模型进行介绍。

§2.1　环境

本文使用WSL2的Arch发行版进行开发，Python版本为3.10.9，Vue.js版本为5.0.8，主要依赖为Element-plus。WSL2是基于Windows开发的Linux子系统，使用子系统可以在性能不受限的情况下使用Windows的大部分计算与存储资源，且传统的虚拟机与双启动系统的开销也不存在。相较于直接使用Windows进行开发，使用WSL进行开发能够使用apt、pacman等包管理器，且Arch还具有详细的官方WIKI与社区支持；而比较使用Linux开发，WSL的Remote连接更加稳定，能够更好地使用WIndows独占的部分软件，例如本文使用的SEERStat数据库官方软件，让开发更加便利。

§2.2　技术栈

§2.2.1　Docker

Docker是一个现今广泛使用的用于开发，运输和运行容器化应用程序的开放平台。使用Docker能够让我们专注于应用程序，而不是花费大量时间调试基础架构。Docker相比较于之前的虚拟机服务，Docker不需要占用多余的磁盘IO，能够有效减少计算资源的消耗。同时，Docker自包含程序依赖，这意味着Python项目的使用中也可以像js前端框架中一样，使用类似package.json的Dockerfile记录所需要的依赖，不同点是Dockerfile不直接写明依赖，而是用requirements.txt等文件来进行存储。这些优势意味着Docker无论在开发还是后期维护中都提供着很强的便携性与健壮性。Docker的核心概念是镜像、容器与仓库。其中镜像不难理解，类似于Linux中的镜像。我们一般使用一个基础镜像，譬如Win10、Linux等，这些基础镜像中包含着能够运行容器的最低限度的底层环境，而我们则基于这个基础镜像编写相关的配置，

比如导入依赖等，然后将这些配置逐层地添加到镜像中，使用Union FS技术对其进行分层与合层记录。镜像的层化技术能让我们具体地分出环境的各层结构，并依据其共享来减少重复镜像的拉取，从而最大化资源的利用率。而容器则包含程序运行需要的一切环境，轻量化地提供可共享可复制地一致服务，容器层的一切修改都不会作用于底层环境，而容器销毁时随着其生命周期的结束所有更改也会消失，从而提供了copy on write的安全特性。仓库类似Github中的仓库，可以使用Docker命令拉取。本文使用Docker搭配代码部署平台后，程序可以在上传到Github仓库之后自动部署程序于服务器上，从而实现方便便携的开发。

§2.2.2 Vue.js

Vue.js是一个面向用户的框架化、结构化的JS前端框架。它建立于HTML、CSS、JavaScript之上，提供声明式和基于组件的编程模型。声明式表明Vue可以使用模板语法动态渲染HTML，让我们可以基于JavaScript中的各个状态动态地描述输出的HTML文本；而组件化意味着Vue的各个模块是可以高度可重用的，Vue SFC将HTML、CSS、JavaScript组合在一起并封装在一个文件中。Vue使用渐进式框架，可以根据需求使用其支持的特性，包括：去构建化的HTML增强、可以在任何页面上嵌入的组件、单页应用程序（SPA）、全栈服务器端渲染（SSR）、静态站点生成（SSG）、可以面向多种应用（包括桌面、移动、WebGL与终端）。

§2.2.2.1 Vite

Vite与Vue-Cli类似，是一个提供项目脚手架与开发服务器的构建工具。不过区别在于Vite并不是构建在Webpack上的，而是使用浏览器中的的ES模块，这让Vite项目提供了很低的延迟和很高的速度，在大型项目中，Vite拥有着远超Vue-Cli的构建与启动速度。在日常使用中，随着项目的不断增大，基于Webpack的Vue-Cli构建速度一般在20秒左右，而相同体量的Vite项目往往恒定在1秒以下，这在需要经常修改的前端项目中会提供很大的便利。这是由于Vite不绑定服务端而是使用浏览器的原生支持。

但是这样的设计也会带来一些问题，Vite的开发环境需要基于现代浏览器，也就是说至少要支持ES2015，在版本较老甚至只使用CommonJS的浏览器中兼容不全，可能会带来一定问题，不过在如今的生产环境中这个问题很少见；同时它暂时也不支持Vue2；脚手架功能中相较于Vue-Cli有一定删节；最后就是开发与构建工具不同可能会导致一些程序页面构建后与开发服务器上的内容不一致。

§2.2.2.2 Element-Plus

Element-Plus是一套基于Vue实现的常用组件库，它提供了较为相当丰富的开源PC组件，这些组件让它成为最富盛名的前端组件库之一，在各类行业的PC前端系统开发中都有着广泛的应用。使用这些封装好的组件能够一定程度上减少开发者自己对于常用组件的再实现与重封装，从而提升开发的速度与稳定性。

Element-Plus的设计原则一共有四条：

- 一致 Consistency: 这表示不但组件的流程与逻辑与生活中使用的一致，而且所有元素和结构亦保证有一致的风格与逻辑
- 反馈 Feedback: 表示用户操作与页面状态都可以让用户感知到不同，从而让用户对网页状态有清晰的认知
- 效率 Efficiency: 说明组件的操作流程直观清晰，用户可以快速而直接地认知到各个结构的用途而不是花额外的时间回忆
- 可控 Controllability: 所有操作都交由用户自身来决策，而且用户可以对已完成或正在进行的操作撤销或终止

Element-Plus受到ES2018以及以上的浏览器支持，而如果要对之前的IE或Chrome版本进行支持，则需要用到Babel或者其他工具进行版本控制。

§2.2.3 Flask

Flask是一个可拓展性极强的、同时极为轻量的后端架构，它的实现是先基于底层的HTTP与Web服务器功能的封装，再使用WSGI(Python Web Server Gateway Interface)来建立Web应用。这样的结构让Flask在传递请求前，需要先将HTTP报文转换为WSGI所需的字典、响应头部的结构体，前者包含请求的全部信息，而后者则是将要调用的函数。而请求信息则由一个显示应用对象处理，在中小型的Flask后端中，大部分的接口都可以定义在这一个对象中。使用显示对象的原因是在Python中，隐式对象只能包含一个实例，所以使用显示对象能够让应用程序集中在一个文件中。这样的设计让小型的Flask服务不会过于臃肿，而且使用者可以使用寥寥几十行代码构造服务，整体结构也相当清晰。而且相当一部分人工智能模型都在Python中有较为简单与便携的实现方式，使用Python来实现后端可以更为简单地嵌入这些模型，让前后端使用的语言数量减少，从而让系统更加健壮。

§2.3 生存分析方法

生存分析是一系列统计分析方法，用于探讨人在特定情况下的，及生存时间的分布由于时间或其他因素的变化趋势。但是生存分析并不仅仅可以用在医学领域，

它还可以在商业等多种环境中使用。比如使用生存分析可以探讨会员、订阅等机制的用户使用情况，并让厂商对于如何留住自身的客户有一定作用。其中生存时间并非单纯表示表示对象的存活时间，在医学数据中，由于数据需要得到用户的允许才能使用，往往数据的获取都是通过随访得到的。这意味着对象可能生存了更长的时间，而我们的数据只能确认对象在一段时间的存活，这种现象被称为数据的右删失。而这需要通过模型的修改或者后期调整来去除。

§2.3.1 KM曲线

Kaplan-Meier曲线是一种基于对时间轴上的生存患者进行累计从而用来进行时向统计的良好方法，它可以用来评估患者群体的健康状况和治疗效果。在生存分析中，我们可以将用户群体按照特征的不同，比如年龄段、肿瘤大小等，分为多个群体，通过比较群体间生存率的区别，我们可以对特征对于患者的影响有较好的初步结论。同时，KM曲线可以用来判断数据是否能使用Cox比例风险模型。

§2.3.2 Cox单/多因素分析

Cox单因素分析一般用来研究单变量中的各个值的关系，它通过类似多项式模型的方法来分析该自变量对因变量的影响，而不纳入其他变量的影响。它主要用来研究该因素对于死亡风险的影响程度。Cox模型定义了风险函数，用来表示一个实体具有相应自变量值时，因变量发生的概率。而Cox多因素得到的p值用来检验该变量对于因变量的影响是否显著，一般p值在0.05以下则表示影响显著，可以纳入。Cox多因素分析则与单因素相仿，研究多个变量对于结果变量的影响程度。

§2.4 其他预后模型

§2.4.1 决策树

决策树是一种在医学中常用的分类回归模型，它使用树形结构表示决策过程。树上的每一个节点都表示特定范围的特征值，而树的每一个叶子节点都对应一个特定的类。决策树的建立过程中，模型首先把全体数据分为大小相近的若干个子集，然后不断在这些子集中选择最佳的子集，让这些子集具有尽量多的相似特征，直到回归达到最大深度停止。决策树可以同时在分类问题与回归问题上使用。

§2.4.2 随机森林

随机森林结合了决策树与随机梯度下降算法，随机地从数据集中抽样，组合成多颗决策树。具体地来说，模型为每棵决策树划分了它使用的数据样本空间，并使

用梯度下降算法让它在样本空间上训练，随后得到一系列决策树组成的森林。由于随机森林分为多颗树训练，在样本数较大的大规模数据集中训练速度快于单棵决策树，同时也可以减小样本的方差。随机森林在NLP、医学诊断学、金融分析学中都有较为广泛的应用。

§2.4.3 聚类+分类

先使用聚类方法得到带标签的新数据，分析聚类得到的患者表型对应集群的生存表现，并根据这个无监督学习的标签分类来预测患者生存率是本文根据文献[2]使用的新方法。在这个模型中，聚类方法的引入让模型的分类更有可解释性，同时不同患者的表型数据虽然在生存率方面会有不同，但是相近的特征依然存在，从而让模型对于相近的数据会有更相似的结果，连续性更好。相比单纯使用二分类模型的概率，该模型也可以使用聚类统计得到的概率作为参考，整体上得到的分类结果更加细致。

第3章　子宫肉瘤预后模型的数据提取

§3.1　数据源介绍

数据提取过程如图3.1所示：

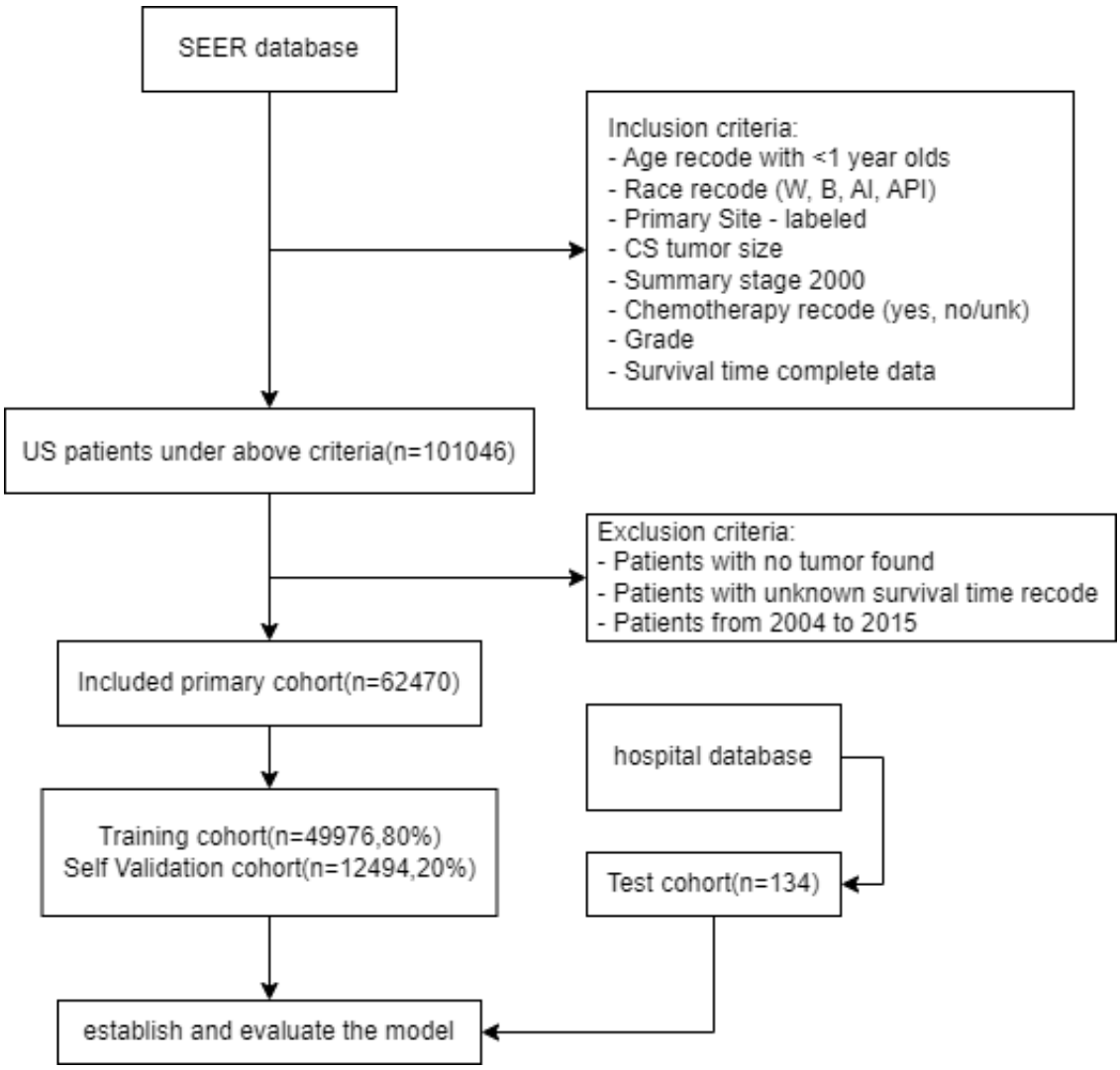


图 3.1　数据提取流程图

由图可见，在SEER数据库中，首先提取从SEER数据库中提取数据，各字段信息将在后面的章节阐释如何筛选得到，最终得到数据101046条。然后通过排除标准，将SEERStat软件中中不能进行排除的部分不需要的数据进行脚本审查，最终分离得到需要的62470条训练用数据。将这些数据进行裁切，最后得到的是49976条训练数据与其附带的自验证数据12494条。

另一方面，由于SEER数据库中的数据存在一定局限性，其只包含被国立医学研究院认可并被SEER接收的数据。这意味着SEER的数据可能由于各地医疗情况与政策差异，不包含某些地区与特定人群的数据，如果文章使用SEER数据库进行验证，无疑会生成部分不够准确的研究内容，甚至影响将来的研究方向。因此，我从医院数据库中进行采集、梳理和标记，最后筛选得到134条验证数据，以供之后验证使用。这些数据可以验证或补充 SEER 数据，以确保研究结果的可靠性和准确性。而且从医院得到的数据也能帮助将来的研究者更好地了解特定疾病的治疗方案与效果，这些信息也会对将来的医学研究与实践具有重要的意义。

因此，使用 SEER 数据和从医院得到的数据都是重要的，以确保医学研究的结果更准确、全面和可靠。两者结合使用可以提高研究结果的可靠性和准确性，并为医学研究和实践提供更准确和有用的信息。

§3.1.1 SEER癌症数据库

SEER（监测、流行病学和最终结果）计划提供大量癌症统计数据，它的目的是通过对癌症数据提供完整的监控数据，减少医疗系统中的癌症负担。SEER数据被认为是美国癌症研究的重要资源之一，因为它包括了来自不同地区的多种癌症类型的发病率和死亡率，以及提供了大量患者的年龄、性别、地理位置和治疗信息等详细信息，同时它也是国际公认的最大最成体系的癌症数据库之一。它起源于1971年美国国家癌症法（NCA）对于建立一个体系化数据库来收集、储存、分析和分发癌症相关数据，以用于支持、预防、诊断和治疗癌症的研究。SEER的病例收集从1973年1月1日开始，在美国的几个地理区间上进行诊断和提取。50年来SEER数据收集范围越来越大，同时内容也不断完整化、规范化，如今收集到的总数据已占美国人口的一半。大量来自不同种族和年龄段的癌症数据为研究者详细分析癌症提供了很大的便利。该数据集的优点是包含了多个癌症类型，且数据集中的样本数量较大，这使得研究人员可以更好地研究不同癌症类型的特点和趋势。

近年来，越来越多的研究人员开始使用该数据集来研究癌症分类和预测，每年大量有关癌症分类和预测的文章在各类期刊上发表，医学人士往往使用列线图等来提供可以在临床上进行定量使用的R语言模型，该数据集也被逐渐用于研究不同癌症类型的生物学特征和发病机制。

§3.1.2 医院验证数据库

医院数据由我在医生老师们的指导下得到，如下图所示：

该数据库提供了患者的病例、检查等数据，在对其中字段内容进行遍历后，我比较了了训练数据中与验证数据中共同存在的一部分内容，并按照第4章分析结果

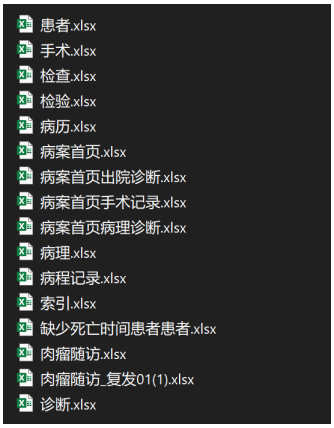


图 3.2　医院原数据图

确定最终纳入的变量。

§3.2　数据字段对应编码与含义

从SEER数据集中，根据分析因素对生存率的相关性选择了下列字段：

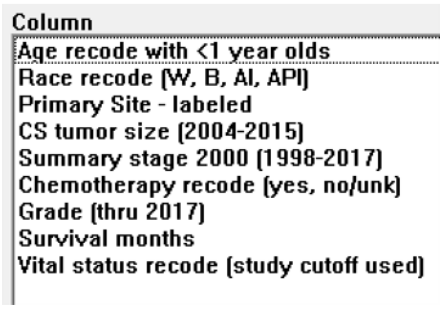


图 3.3　SEERStat中选择的变量数据图

其中方括号内的时间段是由于编码方式改变产生，在不同的时间段有不同的编码方式，所以文章最终取交集2004-2015年范围间。关于字段的解释如下：

- 年龄（例：60-64 years）
- 种族（白人、黑人、美国印第安人/阿拉斯加原住民、亚洲人或太平洋岛民）
- 原发部位（子宫内膜、子宫肌层等）
- 肿瘤大小（最大直径）
- 肿瘤分期（Localized、Regional、Distant）
- 是否化疗（是、否/未知）
- 肿瘤分级（I、II、III、IV）
- 生存时间相关信息

文章中使用到的字段主要如上述列表所示。

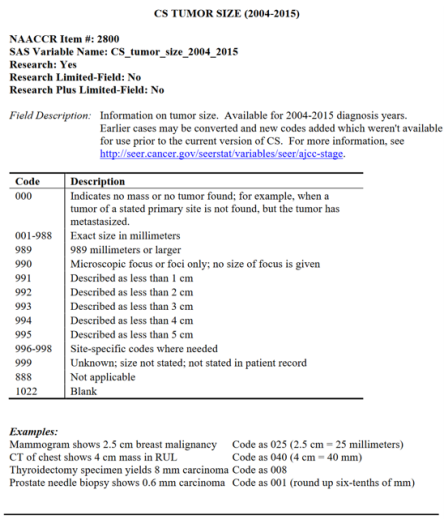
首先，年龄使是以5年为一个单位所表示的，年龄重码变量是基于诊断时的年

龄（单年年龄），使用的分组是由年龄决定的，年龄重新编码变量中使用的分组是由患者数据中的年龄分组决定的。这个重新编码在年龄上有19个年龄组重码变量中有19个年龄组（¡1岁，1-4岁，5-9岁，...，85岁以上）。

第二个字段是种族，在SEER数据库中，主要有六个选项，他们分别是白种人、黑种人、美国印第安人/阿拉斯加原住民、亚洲人或太平洋岛民、其他未说明的（1991年以上）。考虑到在测试数据集中，我们使用的是中国人作为主要的测试患者，一开始我只使用了亚洲人或太平洋岛民的数据，但是在之后的模型建立中，我发现如果只使用亚洲人和太平洋岛民作为训练数据的话模型的效果并不好。可能的原因是亚洲人和太平洋岛民中存在着显出的差异，另一种可能的解释是数据量的不足。考虑到该字段频数的差异我最后使用了白种人黑种人和亚洲人或太平洋岛民作为训练集中开始算的参数选项。

文章使用的第三个字段是原发部位（Primary Site），这个字段主要表示病发的部位。这个字段提供了ICD-O-3规范的主要部位代码和一个描述性的主要部位标签该标签是首选的ICD-O-3加粗的名称，其他部位或子部位包含在代码中，但没有反映在代码中，本文使用的子宫肉瘤数据都包含ICD-O-3标签。编码中可能还包括其他部位或子部位，但没有反映在首选标签中。诊断年份在1992年之前的病例从早期版本转换为了ICD-O-3。这里亦只选取了四个选项，见图3.5中。

第四个字段是肿瘤大小，肿瘤大小指的是肉眼所见肿瘤的最大直径。肿瘤大小适用于2004-2015年的诊断年份。早期的病例部分会被转换，并增加新的编码以加以匹配，这些编码在目前的CS版本之前是不可用的，而且在当前版本的CS之前无法使用的新编码。关于肿瘤大小的详细说明见图3.4。



CS TUMOR SIZE (2004-2015)

NAACCR Item #: 2800
SAS Variable Name: CS_tumor_size_2004_2015
Research: Yes
Research Limited-Field: No
Research Plus Limited-Field: No

Field Description: Information on tumor size. Available for 2004-2015 diagnosis years. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS. For more information, see http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage.

Code	Description
000	Indicates no mass or no tumor found; for example, when a tumor of a stated primary site is not found, but the tumor has metastasized.
001-988	Exact size in millimeters
989	989 millimeters or larger
990	Microscopic focus or foci only; no size of focus is given
991	Described as less than 1 cm
992	Described as less than 2 cm
993	Described as less than 3 cm
994	Described as less than 4 cm
995	Described as less than 5 cm
996-998	Site-specific codes where needed
999	Unknown; size not stated; not stated in patient record
888	Not applicable
1022	Blank

Examples:
Mammogram shows 2.5 cm breast malignancy Code as 025 (2.5 cm = 25 millimeters)
CT of chest shows 4 cm mass in RUL Code as 040 (4 cm = 40 mm)
Thyroidectomy specimen yields 8 mm carcinoma Code as 008
Prostate needle biopsy shows 0.6 mm carcinoma Code as 001 (round up six-tenths of mm)

图 3.4　SEER数据库中CS Tumor Size字段的详细说明图

肿瘤分期是指对恶性肿瘤进行分期分类的过程，它用于描述肿瘤的扩散程度和

位置等特征，以便更好地评估肿瘤的严重程度和治疗前景。肿瘤分期通常由国际癌症联合会制定，其分期系统根据不同的肿瘤类型和特定癌症的组织学和生物学特征而有所不同。SEER主要使用的是TNM分期系统。TNM是一个在国际中广泛应用的为肿瘤进行而实现的系统，在它的分类标准中，主要针对以下几个方面进行分期：肿瘤的大小、淋巴的侵扰情况、肿瘤的转移与否。这里使用的是M标准，肿瘤的转移与否，主要分为三种情况，分别表示局部区域与远端三种类型。其中局部表示肿瘤仅限于原始的发生位置，没有扩散到周围组织或器官。区域表示肿瘤扩散到周围组织或病变器官，但没有扩散到远处的器官或淋巴结。远端则表示肿瘤已经扩散到其他远隔的器官或淋巴结。肿瘤分期表示了肿瘤扩散的进展过程，是对患者生存率有较大影响的一个影响因子。

化疗与否标识了患者是否使用了化疗作为辅助治疗手段。这里要主要的一点是由于患者隐私和机构编码原因，使用的两个编码分别是“是”和“否/未知”，这样的编码会对准确率产生一定的影响。

肿瘤分级是指对恶性肿瘤进行分期的的一种方法，它可以根据肿瘤的大小、形状、密度和边缘等特征来评估肿瘤的恶性程度。肿瘤分级通常由医生进行视觉评估，也可以通过计算机辅助断层扫描（CT）、磁共振成像（MRI）和其他影像学技术来进行辅助评估。这里一个分为四个等级：

- 等级一：这个等级是最低的，在这个等级中，肿瘤的分化情况最低，肿瘤细胞构成的组织和正常工作的组织差别较小，这种情况也常常被书面称为分化良好，肿瘤的恶性化程度最低。
- 等级二：在这个等级中，肿瘤细胞已经开始出现一定的明显不同，这不但体现在肿瘤细胞的生长周期已经与正常细胞有明显的差异，而且肿瘤的形态也与其他组织有不同，这类组织恶性程度已经较高。
- 等级三：这类肿瘤组织等级被称为分化最差的，这意味肿瘤的分化程度较差，大部分肿瘤细胞已经没有正常组织的功能表现了，内部是大量无定形的而无法描述的细胞，这类细胞由于分化程度低，所以增殖速度也最快，不成组织的细胞也最易向周围组织侵扰，所以危险程度也最高。
- 等级四：这个等级表示未分化，这意味着肿瘤的分化程度最低，其中几乎没有分化的细胞，在临床中出现极少。

以上数据中，肿瘤大小、肿瘤分期、肿瘤分级与生存相关内容外的数据都可以直接使用SQL语句或Python脚本简单提取，这里我使用了ipynb文件作为提取方案，以适应研究过程中的大量更改。生存相关数据在数据库中并没有记录，所以我使用脚本筛选了所有提供联系方式的患者的信息，在医院方面进行随访后对得到的数据进行提取。

而肿瘤大小、肿瘤分期与分级数据没有直接字段提供，只能从病理分析中的诊断文本中提取，我首先筛选了其他字段记录完整的患者，并在这些数据的诊断信息中提取。肿瘤大小这里选取的是左附件区和右附件区中的最大肿瘤的最大直径，所以我编写了相应的Python 简单NLP算法，文本的左附件右附件的划分，并在这两段中分别找到用以描述肿瘤的大小。由于还有关于回声区和输卵管长度等的干扰，必须严格找到肿瘤对应的大小。最后在得到的所有肿瘤大小中找到最大值并返回。无匹配内容的描述只能得到空值作为回应，所以最后我对所有数据进行了人工校验。而肿瘤分期与分级同理，先在文本中检索是否有直接指明的描述，如果没有则按照上文中的定义来确定相应的内容。

通过上述的处理，初步筛选得到的300条数据在去除上述内容的缺失后一共是134条测试数据。

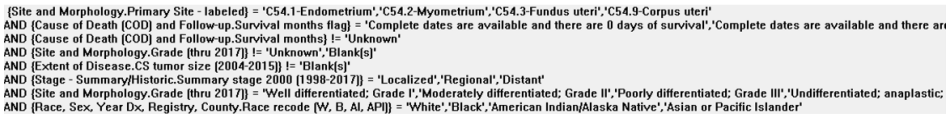


图 3.5　SEERStat中选择的变量排除条件图

第4章　可行性分析与生存分析

§4.1　可行性分析

§4.1.1　技术可行性

通常对于生存分析与建立预后模型使用的是R语言，然后考虑到后期需要将训练得到的模型与后端进行组合，而且R语言的使用人数较少，没有完整的社区资源可供查询，而且为了减少最终系统的复杂度，最后使用了Python作为主要框架的语言。而Python中虽然官方包并没有支持生存分析与模型建立的内容，但是Python有大量第三方包可供调用。在Python中，可以使用的生存分析包的数量没有限制，只要这些包相互兼容，不造成冲突。如lifelines、scikit-survival、survival、pysurvival和scipy。每个包都有自己的优势和劣势，包的选择取决于分析的具体需要和目标。lifelines具有易于使用的API，用于常见的生存分析任务，如Kaplan-Meier估计、Cox比例危害回归和加速失败时间模型。支持随时间变化的协变量、左截断和区间删减。提供可视化工具，如生存曲线、危险函数和累积危险函数。虽然它对灵活的参数化生存模型和对竞争性风险分析的支持有限。但是在目前的标准下其已足够完成分析任务。而相比于lifelines，其他包比如scikit-survival相对于图形化的支持不足，而pysurvival由于复杂的语法和有限的文档，学习曲线可能很陡峭，scipy则ui较为简陋。所以最后选择了lifelines作为主要分析工具。

在后端方面，相比于Java的SpringBoot框架，Flask的环境要求相对较小，也比较容易编写相应的Dockerfile以满足简单的部署。同时考虑到和使用Python的模型结合，如果使用相同语言则可以直接将模型嵌入其中，所以最后使用了Flask作为主要框架。而笔者在大学期间与实习过程中已经积累了相应的在服务器、开发板中部署环境的经历。考虑到经济价格原因，最后的Flask项目先后在阿里云与树莓派开发板上完成过部署，已经确认Flask端的程序具有良好的部署性能与健壮性。而且Flask的良好接口文档也让这项工作有更多的资料可以参考学习。

在前端方面，本文使用的是Vue3作为框架，这里使用的是组合式API开发，相比于Vue2延续而来的选项式API，Vue3的结构更加清晰，使用了少缩进的组合展开方式，可以直接在setup中使用ref、watch等函数来分别管理变量和监控生命周期。这也为开发和后期更改提供了便利。

综上所述，通过分析患者生存情况训练预后模型，并开发使用Vue3-Flask的的管理系统具有技术可行性。

§4.1.2 其他可行性

本系统成本主要包括软件开发成本、版权使用成本、硬件维护成本、数据收集成本。其中软件开发完全由笔者承担，除了占用的时间资源外不消耗其他费用。UI界面中使用的主要是Element-ui提供的图表等资源，MIT协议起源于麻省理工学院，是在所有软件许可中比较宽松的一种，授权人或使用者具有完全的使用、复制、更改、分发等等权力，所以使用相应的资源并不会消耗版权成本。硬件分别使用了阿里云和树莓派用来测试部署效果，阿里云仅仅使用了一个月，成本为12元，而树莓派在之前就已获取，在家运行成本忽略不计。数据来源分别来自SEER数据库与本地医院，SEER数据集中的数据较为完整，笔者通过提交申请后得到授权，随即即可访问相应软件并通过提取得到；而医院数据由于有关患者隐私具体细节不能公开，主要由数据库中进行笔者进行提取，而医院方面配合进行随访，最后由笔者与医院方面共同标记数据得到最后的测试数据。

综上所述，本文进行开发的资源和内容明确，经济成本在可负担范围内，同时使用的材料和内容都不侵犯相关法律条例或其他团体或个人权益，同时时间成本和开发技术栈都进行了有效的权衡，因此该项目可行性得证。

§4.2 需求分析

§4.2.1 目标对象分析

预后模型的主要面向对象是术后的子宫肉瘤患者和相关专业的医生与管理者，考虑到为了避免可能存在的恐慌和信息泄露，该模型及相关系统不会直接向患者或者大众开放，而是用于医疗数字化的应用，比如在医院内网或其他方面搭建。而考虑到使用对象是来自医院的医生、管理员以及后期维护人员，所以一个清晰的UI界面和所见即所得的管理面板是需要的，同时主要的两个用户角色也需要配备相应的权限管理与账户管理。同时为了方便管理员或后台维护人员进行后期的内容增广与技术更新，系统内部同样要配备一定的工单或消息处理与记录功能。从而能让各个角色都能迅捷而高效的完成各自所需要完成的工作。

另外，考虑到数据收集过程到最后模型中预测的不同，同时由于JS到Flask中的表单信息和数据格式需要进行的变化。从前端到后端需要设计一个映射从而不影响数据的准确使用和训练。

§4.2.2 功能需求

考虑到医院、医生多方面的需求，以及患者的实际需求，预后模型有以下几个功能或特点选项：

1. 对于患者的预后数据，模型需要能够输入规定的所有患者数据，包括患者的年龄、肿瘤大小等，并根据这些数据进行推断
2. 预后模型的输出要求可以有多种，比较常用的做法是单单对患者的三年或五年生存率进行判断，如果模型具有概率层，一个概率也将是可以选择的输出，可以用来描述该结果的准确度，以供医生了解结果的准确度。而在后文中，笔者将介绍一种新的方法，使用聚类来将患者分群，此时预后模型的输入相同，而输出则是患者的表征群。这个方法看似只评估了聚类的结果，结果没有之前的有意义。然而，实际上这个结果更加有统计学意义，因为相近的病情状况往往意味着病情的危险程度相近，所以这个结果也具有更好地判断有效性，对于临床使用的可行性也更好。
3. 模型需要分析出各个输入因素的重要性，从而令模型有更好的解释性，对于树型模型需要有图显示其结构。

而一体化平台则需要具有以下几个功能：

1. 一个登录界面，用来进行基础的账号管理，登录时需要提供用户名与密码。后台存储的账号分别分为管理员与普通医生，需要hash加密密码以让用户登录更加安全，管理员具有访问更高级别页面的权限。考虑到信息安全问题不提供直接使用token或其他第三方机构账号登录。
2. 系统首页，首页可以显示用户的头像和登录位置等信息，同时也显示当前用户上传的数据量与预测患者人数，通过对该功能的增广可以让管理员更方便地布置任务，显示其他信息等。
3. 数据表格页面，数据表格可以显示按照上述规范的带标注的详细数据信息，数据实际信息在后台通过端口处理将接收到的文件按照映射表转换为csv格式的编码信息。这样的操作可以明显减少数据文件的体积，减少占用的内存。同时，数据表格页面同时提供了表单功能，从而方便医生录入新的数据并校验数据正确性。
4. 工单记录功能，有关系统本身和待处理的信息都会在这个页面显示，符合权限要求或指定的接收用户能够在这个页面看到这个信息，而如果信息已经被读取则会标为已读。同时还提供了回收站功能来防止用户误删信息。
5. 训练模型页面，训练模型页面提供了直接上传数据文件的功能，上传数据文件后，文件将通过js判断基础格式是否符合要求，如果符合就会被上传到后端中进行模型训练，训练完成后模型文件将会暂存在后端中，同时该页面下的数据表格也会显示当前的模型名称和评分等信息。如果选择不保存或删除模型，后端将会同时删除使用的表和模型文件。
6. 预测数据页面同样提供了表格用来存储历史数据。用户可以选择使用的模型

并通过直接填写页面中的表单并点击预测键即可预测患者的状况，为了方便起见，这里只选取了患者三年死亡与否的结果预后模型。在这里患者的详细信息被隐去，取而代之的是患者的三年生存与否和当前结果准确率的数据。这些历史记录同样有删除功能

7. 权限管理页面，权限管理页面提供了上述各页面的权限更改功能。其中只有管理员可以修改用户角色，而也只有管理员可以添加新的角色并赋予给用户。管理员可以给用户角色设置对每个页面能否访问。

§4.3 变量选择

在研究过程中，笔者首先阅读了一些有关子宫肉瘤列线图与模型相关的文章[3]，在这些文章中，年龄、种族、地区、婚姻状况、肿瘤大小、淋巴结状态、分期、分级、化疗、放疗都是比较常用的数据。然后在对SEER的数据集进行筛选后，我发现主要的库中婚姻状况等参数由于隐私原因被移除，而淋巴结的信息符合标准的过少，如果使用则会大大削减符合要求的数据数量，从而让训练集变得很小，所以考虑到这点后没有纳入。与医院老师们沟通后，笔者发现分期中的一些描述包含了淋巴结是否被侵犯的状况，而目前医院内主要使用化疗作为主要手段，放疗的数据较少所以没有考虑纳入，所以最后选择了选择了当前的变量。

§4.4 生存分析

§4.4.1 数据概况

从SEER数据集中采集到的数据一共有62470条，患者的年龄分布如下图所示：

由于基数较大，同时表格中的数值取区间中的最低值，表中的频数对其进行了平均缩减以方便显示，可以看到表格中用户的年龄主要集中在60-70之间，这说明了患者的年龄越大，越有可能患子宫肉瘤，如果按照三分位数进行划分，我们可以得到子宫肉瘤的主要发病人群集中在45岁到65岁之间。

而对于另外一个数值变量，肿瘤大小的分布如下图所示：

按照上述表格同样可以得到肿瘤大小的大致分布，在表格中可以看到，肿瘤大小的中间值在50mm左右，肿瘤大小为0mm代表1mm以下或较难以衡量的数据，而肿瘤大小在到达50mm的峰值后则呈反曲线状态减少，直到到达最低点。肿瘤大小的分布较为复杂，因为这个原因所以我们也很难为其划分一个清晰的界限来划分患者，如果直接划分很可能会损失一部分信息。这里先使用50mm为分界点进行划分，后期再是用其他方法进行比较。

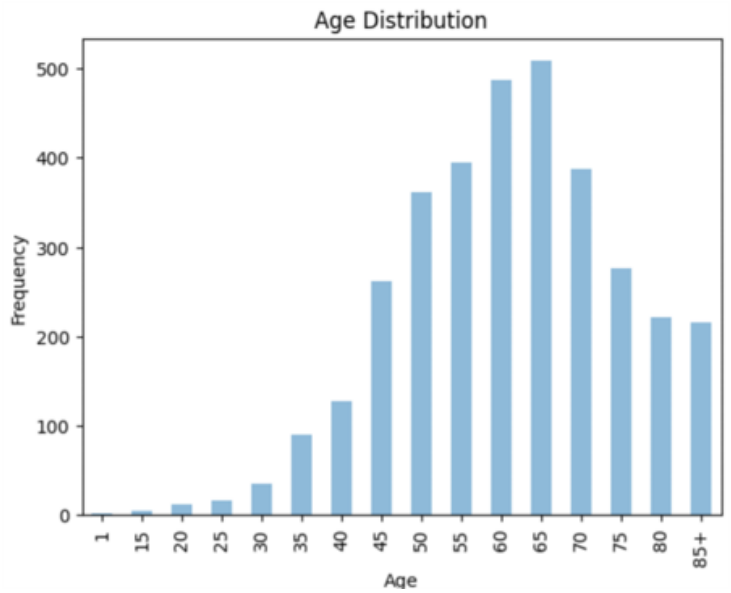


图 4.1　SEER数据集中患者年龄分布图

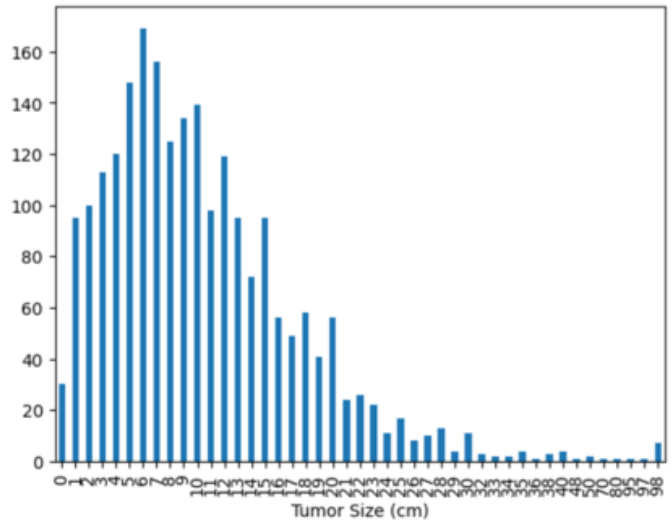


图 4.2　SEER数据集中患者肿瘤大小分布图

§4.4.2 Kaplan-Meier曲线绘制

由于下面使用的Cox比例风险模型需要数据满足以下假设：

1. 风险比值的对数与协变量之间呈线性关系
2. 风险比值的对数与时间无关

上面的风险比值的对数与协变量之间呈线性关系假设需要使用Kaplan-Meier曲线进行验证。具体方法是：对于同一自变量，如果本身为分类型变量直接使用，如果为连续或数值型变量，则把变量分为多个区间作为不同值，观察曲线间状态，如果曲线存在交叉则说明其不满足假设，反之则说明风险比值和协变量间存在线性关系，满足条件。

同时，KM曲线不但可以验证后续的假设是否成立，同时也能对患者的生存状况做更细致的分析。KM曲线的原理是首先计算在某一时期存活的病人能活到下一时期的概率，然后将存活概率逐一相乘，得到相应时期的存活率。这种方法可以非常清晰地看到两组患者在不同时间的生存率差异。

由于篇幅问题，这里仅列举两张比较典型的进行分析说明：

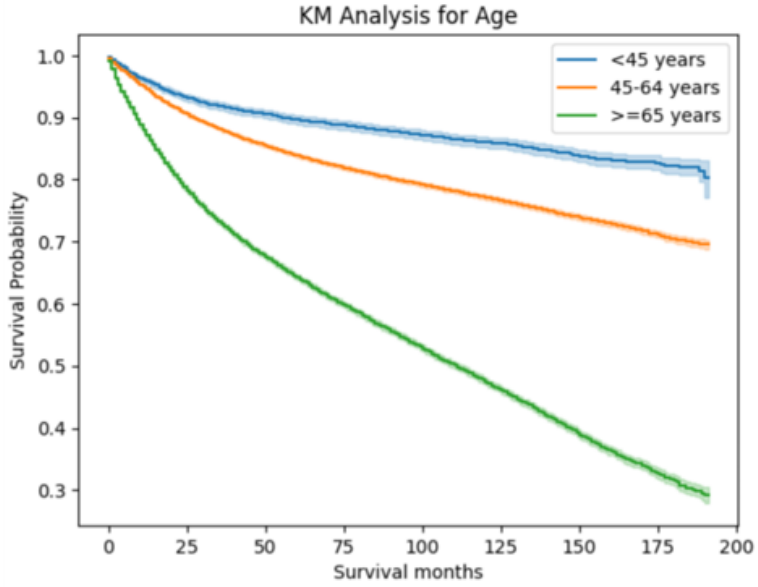


图 4.3 SEER数据集中患者年龄KM曲线图

从图4.3中可以看到，45岁以下的患者在时间变化下衰减最少，而45-64的次之，65以上减少最快，而这些曲线都不交叉，置信区间也较小。这说明了患者的年龄是一个重要影响因素，而患者的年龄越小，生存率越高。

从图4.4中可以看到，四个原发部位的患者生存率从高到低的变化，从置信区间的的范围可以看到四者的数据量也存在差异。同时这些组间也存在线性比例关系，所以满足假设。

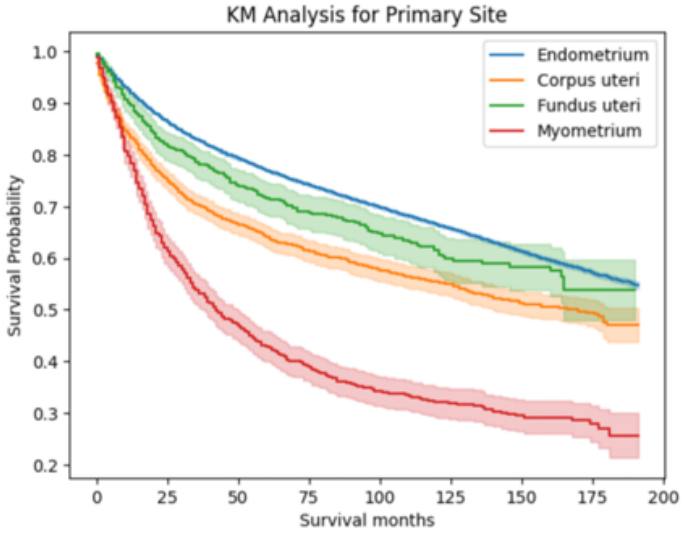


图 4.4　SEER数据集中患者原发部位KM曲线图

§4.4.3　Cox单因素分析

Cox单因素分析是一种基于Cox比例风险模型的统计分析方法，它可以用于研究变量对于结果的影响。本文使用年龄、种族、原发部位、肿瘤大小、分期、化疗情况、分级建立Cox模型，并使用SEER数据集中的数据编码为合适格式，用编码后的结果训练。下列的表4.1是单因素模型的Cox分析结果。

coef是Cox中的回归系数beta，因此exp(coef)是Cox比例风险模型中的概率风险比。从结果中可以看到，计算得到的P值是用来评估同一变量的协变量对因变量的影响是否显著的参数。这里使用变量中的第一个协变量作为参照，分析其他协变量是与对照变量存在显著差异。一般在0.05以下则说明变量对于结果的影响较为显著，可以看到表中的结果都在该值之下，所以说明每个变量的不同值对于计算结果的影响较为显著，同时也验证了在Kaplan-Meier曲线中观察到的结果。

§4.4.4　Cox多因素分析

Cox多因素分析也是一种基于Cox比例风险模型的统计分析方法，它可以用于研究多个不同的变量对于结果的影响。下列模型同样使用年龄、种族、原发部位、肿瘤大小、分期、化疗情况、分级建立Cox模型，并使用SEER数据集中的数据编码为合适格式，用编码后的结果训练。不同的是单因素将变量的每个值作为参数导入，而且同时只导入一个变量，而多因素分析则直接导入有不同值e多个变量。下列的表4.2是多因素模型的Cox分析结果。

从结果中可以看到，计算得到的P值是用来评估多个变量对因变量的影响是否显著的参数。这里由于不是同一变量，所以不需要像单变量分析一样使用变量中的第一个协变量作为参照，分析其他协变量是与对照变量存在显著差异，相反，由于

多个变量互不交叉，其是并不需要设置一个变量作为参照组。一般在0.05以下则说明变量对于结果的影响较为显著，可以看到表中的结果都在该值之下，所以说明每个变量对于计算结果的影响较为显著。这是在Kaplan-Meier曲线阶段无法观察到的多变量影响，这说明不但单个变量的不同值对于结果有比较大的影响，同时这个结果也可以在多变量模型中观察到，进一步验证了所选取的变量都有纳入价值。

§4.4.5 生存分析结论

在上述对于患者数据库内数据分布的分析后，本文分别进行了Kaplan-Meier曲线的绘制和Cox单变量和多变量分析。其中对于数据的分布分析让我们对于数据库中的人群特征有了一定的了解，同时也为如何对变量类型为数值变量的变量进行了一定的划分，从而方便后续在Cox模型中导入并训练。而Kaplan-Meier曲线的绘制有两个目的，第一个目的是可以初步对于不同变量对结果的影响，从而让我们能够初步对于是否纳入该变量有一定的预判，在之前的工作中，笔者对于如何纳入变量有了较多考量，其中首先从论文中选择了一些常用的、对结果影响较大的变量，同时也选择了一些其他的未纳入过的变量，而这些变量中的一部分便因为KM曲线结果不符合模型标准或在这之前便发现变量内容与其他变量具有重复或冲突，所以最后没有纳入。而如果在这个阶段提前使用Kaplan-Meier曲线进行分析的话可能就能避免之前的重复劳动。Cox单因素模型分析选择了每个变量的不同值，将它们进行了拆分为二元变量并堆叠为组的过程，从而导入Cox单因素模型进行分析，在这个过程中，可以得到之前使用Kaplan-Meier曲线不能看到的的精确数值表示，这样便可以让我们对于数据是否能够纳入得到一个更明确地标准，而不是对于曲线进行经验判断。而多因素分析则解构了多个变量，分期它们结合起来对于结果的影响是否显著，这里与之前不同，如果只有单个变量只需要使用 log-rank 检验来确定因素之间的显著性关系即可，而多变量则需要使用比例风险回归模型才能确认影响，如果某一变量在 log-rank 检验中显示影响较大，在多变量回归中可能会因为其他变量占比较大从而反而显示不出较好的影响。而当前结果表明，目前采用的变量都对于结果有较为显著的影响，在之后的模型纳入与训练中可以作为参数使用。

表 4.1　预测子宫肉瘤患者总生存率的单变量Cox分析

variable	exp(coef)	P value
Age		
<45years		
45-64years	1.68	<0.005
≥65years	4.35	<0.005
Race		
White		
Black	0.56	<0.005
Asian or Pacific Islander	0.47	<0.005
Primary Site		
C54.1-Endometrium		
C54.9-Corpus uteri	1.67	<0.005
C54.3-Fundus uteri	1.43	<0.005
C54.2-Myometrium	3.24	<0.005
Tumor Size		
<50mm		
≥50mm	2.36	<0.005
Stage		
Localized		
Regional	2.72	<0.005
Distant	10.94	<0.005
Chemotherapy		
Yes		
No/Unknown	0.41	<0.005
Grade		
Well differentiated; Grade I		
Moderately differentiated; Grade II	1.84	<0.005
Poorly differentiated; Grade III	4.93	<0.005
Undifferentiated; anaplastic; Grade IV	6.95	<0.005

表 4.2　预测子宫肉瘤患者总生存率的多变量Cox分析

variable	exp(coef)	P value
Age	2.23	<0.005
Race	0.82	<0.005
Primary Site	1.17	<0.005
Tumor Size	1.52	<0.005
Stage	2.47	<0.005
Chemotherapy	1.29	<0.005
Grade	1.56	<0.005

第5章　预后模型的设计与实现

§5.1　模型需求与衡量标准分析

按照实际需求，预后模型的主要目的是将不同类型的预后群体区分。与其他领域使用模型准确率进行模型的评估不同，由于医学模型的建立初衷是评估患者的实际表征的区别。按照常用模型的分类标准[4]，现在学术界经常使用的是将生存数据之外的变量作为参数输入，并使用预测3年或5年生存结果或概率估计作为输出，这是一个基于时间-事件的回归问题。这种方法得到的决策树模型能够按照决策树结构中的先后顺序来判断特征重要性，而且模型能够按照节点导出一张树形图。虽然这个方法能够得到一些关于重要性的内容和一张树形图，但是考虑到这样的结果可解释性较差，所以本文在这里引入了一个新的方法。聚合层次聚类方法常常被用来将患者的临床表现分组，按照患者的临床表征相似度来区分患者的不同风险层级，从而能够让聚类方法中的区分效果更好。这里为了维持患者的分类效果，让层级效果更高，所以有必要对于患者的表征分类数量进行一定的评估，但是考虑到临床要求，分类数如果过少或者过多依然不能很好地区分患者，因为如果层次过少无法很好地细分患者的风险，如果过多则对于日常使用有一定影响，同时聚类效果也会较差，所以这里需要按照实际临床需求和实际效果进行权衡。如果能够对于患者进行分群，临床上能够对于患者的不同特征还得更好的数据，从而能够更好地评估每个患者的类型，如果将来的临床治疗方法发展，也能更好地对这些患者的实际情况进行区分，因此也能让这些患者群体的患者使用不同的治疗方法，对症治疗后的评估也能让之后的数据与之前的数据形成对比，发现不同患者群的预后效果变化，这类数据分析也能帮助患者更好地获得符合自己的治疗效果，同时也能让患者对自己的病情特征有更好的了解。然而，虽然无监督方法是探寻患者表型的有效工具，与有监督模型相比，无监督模型在对新数据的应用上有局限。这具体表现在：患者的无监督模型在对于一个数据集训练后如果得到了表型，那么这个模型虽然能够区分出多个表型，但是对于每个表型需要人工进行挑选，一般会呈现为低中高等多个等级，而这些表征等级的不同特征往往是比较独一无二的。如果要对新的数据重新进行训练，那么原先的表型很有可能会与新的表型有一些不同，对于K-means聚类，这就体现在每个簇的质心就算在新的数据上表现类似，但是依然不会完全一致，这造成了聚类方法在新的训练后依然需要人工区分新的表型。那么能否完全使用有监督模型替代呢？答案是否定的，有监督模型无法直接仅凭自身方法总结出这些特征。综上所述，结合无监督与有监督ML方法是一个可能的解决方法[2]。

考虑到上述需求，能否区分患者的实际表现差异更为重要，所以预后模型的价值一般使用CI进行验证，好的预后模型CI值较高。CI即为C-Index，是用来评估模型预测能力的指标，它主要计算了模型预测值与真实之间的区分度。在肿瘤患者的实际预后精确的衡量中，我们很多时候是从两个层面进行评估，其中一个是使用拟合优度检验指标进行衡量，这个标准主要用于衡量分类结果中的各个分类的频数的期望，从总体分布状况上。不过在医学临床应用角度中，一个更重要的指标是对模型精度的衡量，即使用C-index来衡量预测结果的准确性。相比于使用常见模型使用的均方差等，更常用的便是使用上文提及的C-Index进行衡量，计算结果发生概率和实际概率的一致性是我们主要的目的。C-index的作用便是预测预期结果和实际输出相一致的概率，在二分类问题中，正类结果大于负类结果即可累计从而计算概率。C-index通常先对数据进行预处理，将使用的验证数据集中的实例两两分为许多对子，这些对子中不是两两配对即可，而是要生成所有的对子组合，所以如果验证数据集中一共有n条样本，那么最后生成的对子数即为C_n^2对。将这些对子中存在没有到达观察终点的无效样本的对子排除后，剩余的即为有效对子。在对子中，如果预测结果都与预期结果相一致，那么这些对子被称为一致对子，统计对子的数量。根据上述过程，设样本数为n，无效对子数为N，有效对子数为A，可以得到以下公式：

$$CI = \frac{A}{C_n^2 - N}$$

考虑到sklearn中的相应指标是ACC，它的计算方法为：

$$ACC = \frac{TP + TN}{TP + PN + FP + FN}$$

而没有直接对于CI的实现，所以这里笔者使用了ROC曲线下面积AUC进行等效替换二分类模型中的CI，这是因为在二分类模型中AUC和CI是相同的，而多分类模型中的则调用其他sklearn之外的专用库函数实现。

§5.2 使用分类模型建模

§5.2.1 决策树

决策树使用树形结构来进行分类，通常决策树一般在分类患者的具体病情中使用，而考虑到本研究指针对子宫肉瘤的患者，而如果用来分类原发部位则不太符合临床标准，因为患者的具体的原发部位使用核磁共振或超声图像进行诊断会更加精确而直接，同时考虑到对患者进行诊断本身就是医生的工作，如果用模型进行分类

不但会与医院本身的工作流程重复，数据内容也不支持建模，同时模型的准确率也很少会高于人工诊断结果，所以不需要使用决策树来对原发部位等信息进行分类。所以这里则直接用决策树来进行患者三年生存概率的二分类模型建立。

决策树选择好特征后，根节点开始是的非叶子节点分别是决策树的树中的特征范围，而决策树中的叶子节点则表示了最终的分类结果。决策树使用了基尼系数来衡量特征空间的有序性，决策树通过改变数据划分来使得信息增益最大，即使得最终的基尼系数最大。基尼系数。

§5.2.1.1 决策树中的特征重要性分析

这里使用的输入数据中，其他输入变量如章4中分析得到，分别为年龄、种族、原发部位、肿瘤大小、分期、化疗情况、分级，考虑到年龄为区间分类变量，所以需要对其进行编码化，按照顺序分别将其编码为0,1,2,3,4...等，而训练的分类目标则是三年死亡率。经过训练后，可以得到如下表5.1的特征重要性表：

表 5.1 决策树得到的特征重要性

Age	Race	PS	TS	Stage	Chemo	Grade
0.0210	0.0028	0.0035	0.0037	0.0295	0.0031	0.053
high	low	low	low	high	low	high

根据上表可以分析得出，年龄、分期、分级相对于其他因素重要性较高，而种族、原发部位、肿瘤大小、是否化疗则相对较低。这些原因可以得到一定解释。

患者的年龄显然是一个与生存率相关的因素，年龄较大的患者免疫系统较为低下，他们的身体机能也往往较弱、呈现退行状态，而因此患者自身对于癌细胞的抵御作用较低，患者同时也有一些并发疾病，比如心血管疾病等，这些疾病让患者更有可能在治疗过程中分身乏术，很多时候死亡原因也包括因为癌症的并发症而死亡，这些因素导致患者的年龄对于死亡率有很高的影响。同时年龄较高的患者也同时是癌症的高发群体，他们更易因为致癌因素的积累等原因患癌症，同时他们也更有可能将癌症误认为老年化后自然而然得到的疾病，在数据中，可以分析得到中老年子宫肉瘤患者分期和进展程度往往明显高于青年患者，这也解释了为何老年患者的生存率明显低于青年患者。

而分期则是癌症进展的一个重要指标，如果癌症的扩散程度较高，那么患者的治疗便需要更为激进，使用一些对于患者身体有大量伤害的疗法，比如化疗和放疗等等。同时，如果患者的分期中，肿瘤细胞扩散到了周边组织，特别是淋巴细胞的

侵扰，会让手术和化疗手段更难清除肿瘤细胞，而癌症细胞的扩散也就意味着复发的可能性更高，所以患者的死亡率也更高。

对于分级而言，它衡量了肿瘤细胞和它的来源组织的结构学相似度，肿瘤的细胞如果相似度高，那么它的分化程度也较高，这说明了癌症细胞由于突变导致的碱基对缺失等的变化较小，这类细胞在微观层面上和周边的原发细胞的差异较小，分化程度高。而因为高分化的细胞往往因为其组成的组织学结构较为稳定，往往嵌合为较为有序的结构，这也表明其继续分裂生长的可能性较低，速度较慢。而如果分化程度较低，甚至为未分化细胞，那么肿瘤细胞的形态上往往是无序而呈现为离散结构。一般而言，越是未分化的细胞往往更容易以较快的速度增值，这些细胞的组织学特征也让它们往往呈现为较小而不定形状态，所以也容易通过血液或组织液等途径向外扩散，所以分化较低的恶性肿瘤细胞往往意味着肿瘤本身是恶性的，对人体的威胁也较大，分级较高的患者如果没有通过及时检查，其恶性肿瘤的扩散速度相对于良性肿瘤一般会更快，这也体现了分级与分期的相关性。

虽然分期分级有一定关联，然而其却不能完全相互替代，这是因为同样分级的患者，按照发现的时间和病情进展情况会有不同的分期作为体现，这是因为分级较高的恶性肿瘤如果发现及时未扩散，那么相比于恶性程度较低但是扩散的肿瘤会更好处理，所以两者之间不会体现替代关系。同时，如果相同分期的患者之间的分级有区别，那么分级较高的患者之后的扩散会更难抑制，同时药物对于不同分级的细胞的作用也会有一定区别。分期分级作为两个衡量指标，一个体现了癌症的进展状况，一个体现了癌症的微观恶性程度。前者是癌症对于患者的宏观影响，后者体现了癌症的特性同时，也表现了癌症的之后的进展速率。所以两者相辅相成不可分割。

分析了上述影响因素较大的特征，接下来是影响较低的特征，分别是种族、原发部位、肿瘤大小、是否化疗。

其中种族影响因数较小的原因显而易见，首先数据集中采用的是美国SEER数据库的数据，其中数据来源近似，医疗水平也较为平均，而同一国家的生活方式，环境区别整体水平也差异不大，趋于同质化。如果数据来源于不同国家，可以预见的的由于不用医疗水平的差异，来自不同国家的不同人种的数据由于环境、生活方式、治疗能力的差别可能会有较大的差异。同时这个特征也体现了不同种族的人在对于子宫肉瘤的死亡率虽然存在差异，但是相比于上述的重要因素并不是主体影响因素。

原发部位体现了患者的肿瘤从哪里的组织开始生长，对于分化较高的肿瘤，肿瘤细胞本身的结构和性状会更接近原先组织，所以原先组织的生命周期特点会对肿瘤之后发展的情况有一定影响，比如卵巢原发的肿瘤相比输卵管肿瘤在相近的分化

情况下可能生长速度会更快。同时，原本组织的形态特征和功能特征也会对患者的身体情况有一定影响，性腺的激素相关细胞较多，所以如果肿瘤细胞继承了原先正常细胞的激素分泌功能，患者的激素水平很可能也会受到相应的影响，而输卵管的病变如果长期影响到了输卵管的正常功能，也可能会导致卵巢等组织的功能受到影响，可能会导致患者需要切除正常组织，这也会对患者的预后结果产生较大的影响。而这个影响因素体现为较低水平的原因可能是这些影响在经过手术和相应治疗手段后比较容易去除，同时医学界也有相应的成体系的治疗方法应对，所以对于患者的预后影响较小。

考虑到子宫肉瘤对周边组织的侵扰性较强，且肿瘤切除后对于患者的影响相比于一些其他部位的肿瘤，比如脑瘤等风险更小。这是因为患者腹腔中的空间较大，较大的肿瘤对于患者的压迫作用较小，手术切除的风险相对而言也不是特别高。同时，子宫肉瘤往往会因为其侵扰性出现多个肿瘤，而肿瘤大小只能反映其中最大的一个肿瘤大小，其他肿瘤并不在考虑范围之内。较大的肿瘤在临床上有多种方法，包括分段切除、淋巴结清扫等，这些方法可以有效降低肿瘤的转移和复发概率。[5]所以肿瘤的大小虽然也是一个重要的特征指标，但是重要性次之。

而是否化疗对于肿瘤的影响较小是因为化疗与否是治疗手段的一部分，一般只有肿瘤进展较高的患者需要进行化疗进行术后辅助治疗。而SEER数据中，化疗数据由于患者隐私问题，所以不化疗和不知道是否化疗的数据被归为一类，这也导致了化疗对于患者的影响因数可能被削减了。

§5.2.1.2 决策树树形结构一览

考虑到实际树形结构过大，难以在页面中完整显示，所以图5.1仅仅列出了一部分树形结构：

由于特征过多，决策树使用了PCA降维来进行主成分分析，通过降维变量来提取数据的主成分向量来对数据空间进行优化。可以看到图中树的深度为15层，基尼系数从0.449开始逐层下降，最终的分类结果则在叶子节点中体现。

§5.2.2 随机森林

随机森林是一种基于决策树实现的集成学习ML算法，它通过建立多个具有不同树核特征的随机决策树构建，树核特征是指这些树具有不同的深度和节点分布。按照数据中的样本权重，各个决策树不断优化其树结构和表现，最后随机森林选取其中效果最好的一类决策树，并利用它作为最终的分类模型。随机森林的生成过程相比于决策树有一定的随机性，同时训练集中的重复数据构成的不用样本权重分布也会让随机森林更好地把握数据之间的关联。由于使用随机方法构建多棵决策树，

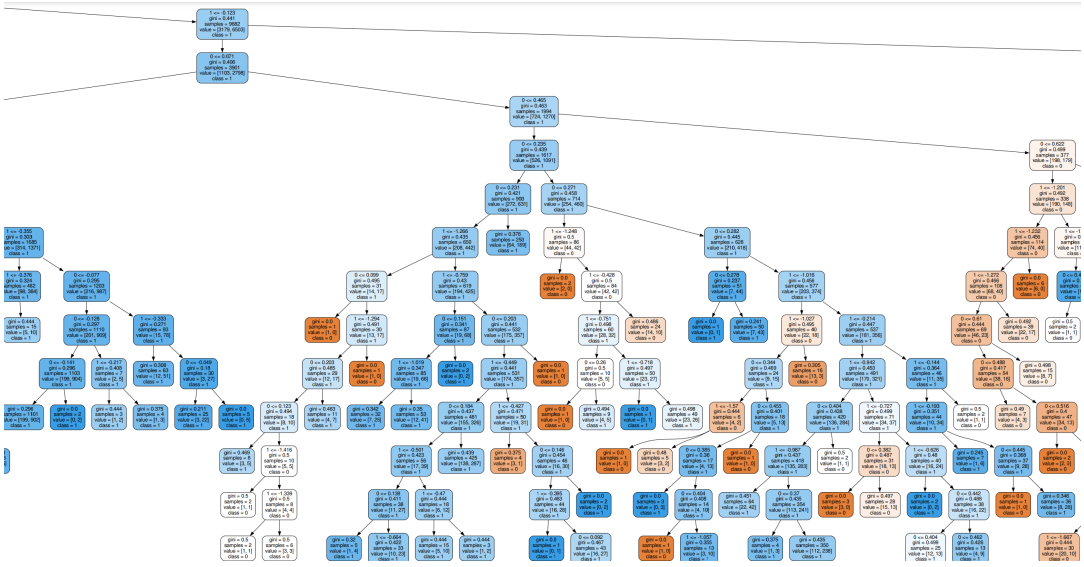


图 5.1　决策树部分结构一览

在数据量有限的训练数据中，相比于决策树，随机森林有更好的泛化效果和鲁棒性，同时它也能更好地分块处理大量数据，因为大量数据中的重复样本分布能让随机森林更好地发现数据的重要程度的不同，拟合效果和训练速度也更好和更快。

§5.2.3　Cox比例风险模型

Cox比例风险模型经过章4的分析已经验证了它的使用前提假设在当前数据中是满足的。该模型主要基于使用风险函数表示：

$$h(t) = h_0(t) \cdot exp(b_1 \cdot 1 + b_2 \cdot 2 + ... + b_p \cdot p)$$

风险函数表示患者在时间t发生死亡事件的风险。

§5.2.4　模型效果评估

经过上述三个模型的训练，表5.2表示了三个模型的训练效果：

表 5.2　决策树得到的特征重要性

模型	训练集C-Index	测试集C-Index
决策树	0.75	0.76
随机森林	0.79	0.78
Cox比例风险模型	0.80	0.79

从图中可以得到一些信息，决策树的拟合能力稍弱于随机森林，而由于数据量较大，所以两者在测试集中的表现差异不大，而Cox比例风险模型由于使用了风险函数进行评估，而且结果输出的是指定时间而非对于患者的特定时间的死亡风险进行评估，所以相对于对于三年死亡训练得到的二分类模型有更强的表现力，不是单单的二元分类也让它的CI指数大于二分类的前两者，这是由于一些处于边界情况的患者较难进行分类的缘故。

§5.3 使用聚类与分类模型结合建模

§5.3.1 聚类

§5.3.1.1 聚类方法与聚类数量确定

这里分别使用K-means和层次聚类方法对数据进行聚类，轮廓系数如表5.3所示：

表 5.3 不同聚类方法的轮廓系数

方法	2 clusters	3 clusters	4 clusters
K-means聚类	0.356	0.239	0.247
层次聚类	0.378	0.337	0.273

其中轮廓系数是用来衡量聚类后的簇间轮廓清晰与否的指标，如果一个采样点和其所在的簇内的其他元素较为贴合，且与簇外的元素紧密程度较低，那么则可以说该次聚类的效果较好。从上图可以看到，层次聚类在当前数据集中的聚类效果好于K-means聚类，而且轮廓系数上聚类的目标类数越多，轮廓系数越低。这是因为患者的特征各有不用，在样本空间上较为弥散导致的。虽然分两类可以得到较好的聚类效果，然而两类的临床作用却较低，所以权衡后，本文使用了三类中效果表现最好的层次聚类。

§5.3.1.2 聚类效果评估

下面的图5.2和图5.3是使用Kaplan-Meier曲线对聚类方法进行的效果检验：

可以看到图5.3中的三类的生存率曲线存在明显的差异，这表明了就算没有将患者的生存情况纳入作为因素，仅仅凭借聚类方法生成的患者表征群仍然具有一定的生存率差异，而这差异则可明显地将患者分为三个风险等级。

在图5.4中，我们可以看到在三个患者表征群中，对于连续变量肿瘤大小而言，三个簇间区分度较好，而对于分类变量而言，三个分类则都具有不同的各个分类值

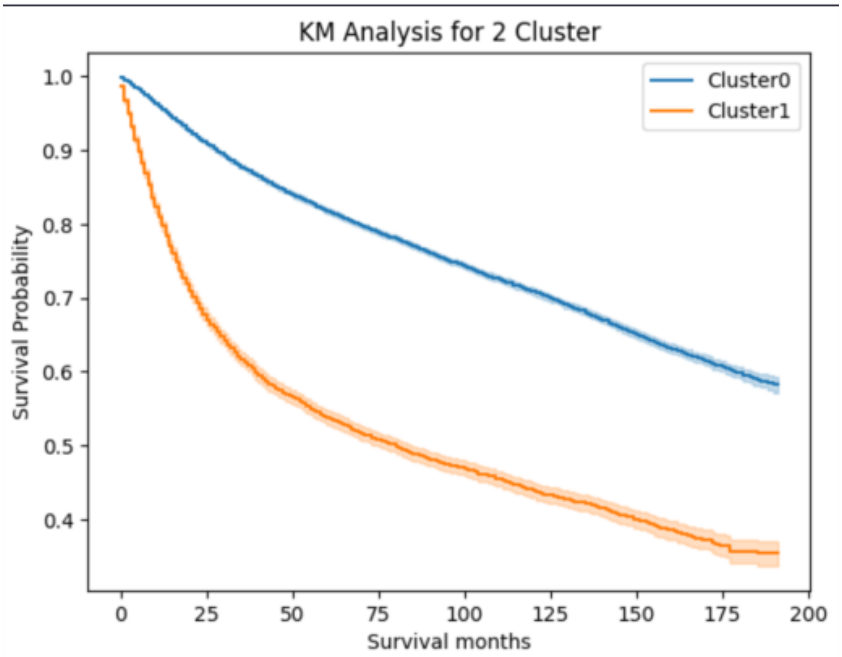


图 5.2　层次聚类分为两类时的Kaplan-Meier曲线

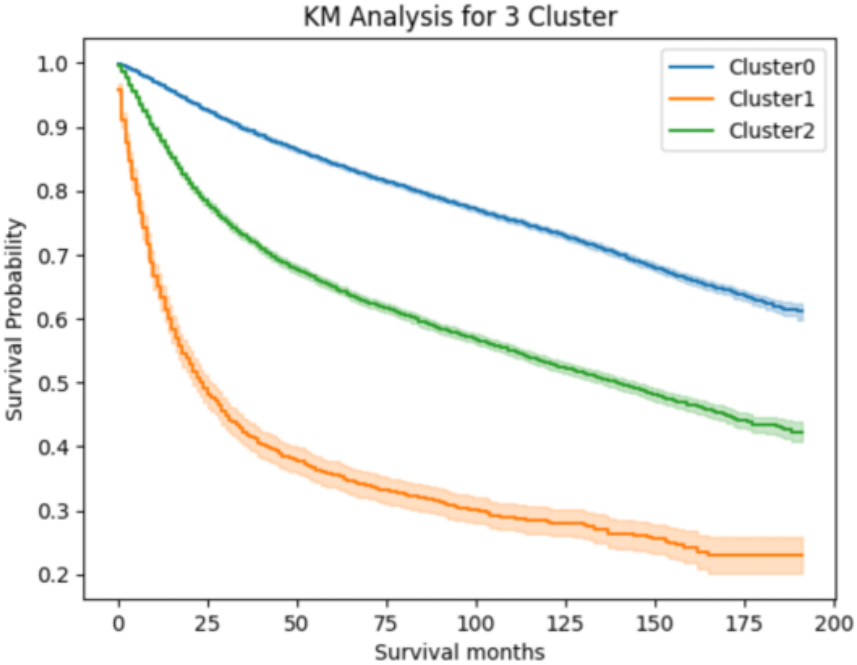


图 5.3　层次聚类分为三类时的Kaplan-Meier曲线

的比例，区别是三者的比例分布有差异，这可能是考虑到了综合不同因素的原因。

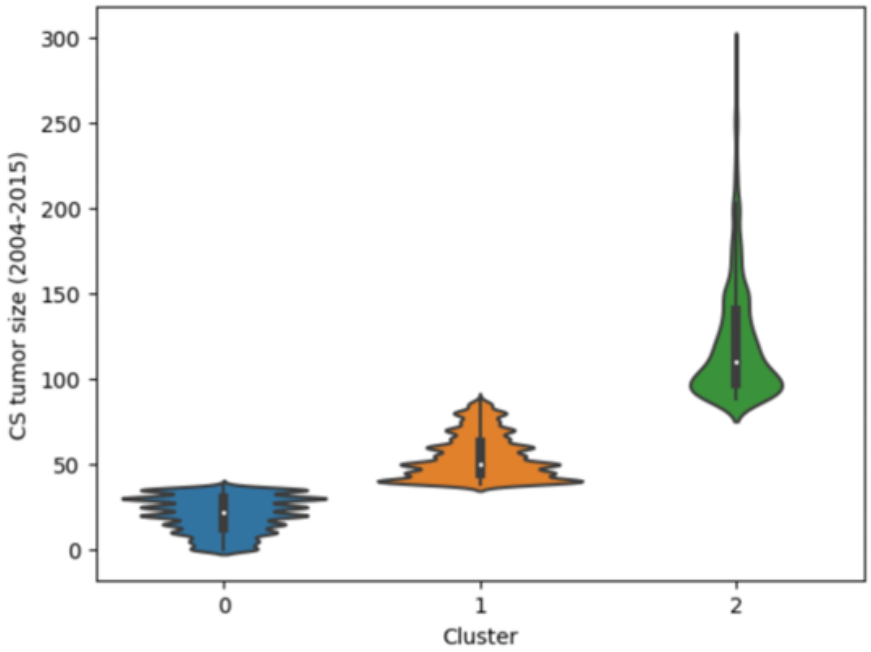


图 5.4　层次聚类分为三类时的肿瘤大小变量小提琴图

总体而言，层次聚类方法能够将患者分为三个生存率有明显差异而不交叉的簇，而对于其中的连续变量也有较好的划分性能，具有一定的参考使用价值。

§5.3.1.3　模型效果评估

在将聚类标签作为分类目标，在训练集和测试集中分别标记，并使用随机森林模型训练后，该模型的训练集CI为0.97，测试集CI为0.95。可以清晰的得到这个模型具有很好的分类效果，原因是分类的目标本身就是聚类得到的空间点位相近的簇，分类模型对于这种比较清晰明确的区间分类拟合效果很好。

§5.4　预后模型总结

本章主要介绍了两类分类模型的实现，其一是基于预测患者的三年死亡与否和死亡时间的分类模型，在这类模型中，本文直接使用监督学习方法进行预测，决策树和随机森林的结果比较相近，两者接近于使用Cox比例风险模型进行分类的结果。而第二类则是使用聚类方法配合监督学习进行的患者表征分类，这类方法不直接得到患者的具体死亡情况，而是根据患者的病理特征对患者群体进行分类，得到多个死亡风险有差异的表征群（簇），这种方法具有较好的临床参考价值和极高的一致性参数（CI），但是考虑到这种方法的目标并非直接得到死亡率，而不同群体之间也不存在百分百的确定生存或死亡，其实际的准确率达不到结果的0.95，但是考虑

到实际患者的生存率取决于多方面影响，并没有与病理得到的参数强相关，进行死亡率分类的模型最后的CI性能往往被限制在0.8左右，所以这种方法为如何相关研究提供了一项新的思路，具有一定的参考价值。

致　谢

衷心感谢导师沈文枫教授对本人的精心指导。同时感谢同济医院袁素珍老师、汪雯雯老师对工作的支持和上海大学开源社区提供的Latex模板。

参考文献

[1] 中国医师协会微无创医学专业委员会妇科肿瘤专业委员会(学组)中国优生科学协会生殖道疾病诊治分会, 中国优生科学协会肿瘤生殖学分会. 子宫癌肉瘤诊治中国专家共识(2020年版) [J]. 中国癌症防治杂志, 2020, 12 (6): 599–605.

[2] Zhou X, Nakamura K, Sahara N, et al. Exploring and Identifying Prognostic Phenotypes of Patients with Heart Failure Guided by Explainable Machine Learning [J/OL]. Life, 2022, 12 (6): 2–5. https://www.mdpi.com/2075-1729/12/6/776.

[3] Song Z, Wang Y, Zhang D, et al. A Novel Tool to Predict Early Death in Uterine Sarcoma Patients: A Surveillance, Epidemiology, and End Results-Based Study [J]. Frontiers in Oncology, 2020, 10: 3–6.

[4] Du M, Haag D G, Lynch J W, et al. Comparison of the Tree-Based Machine Learning Algorithms to Cox Regression in Predicting the Survival of Oral and Pharyngeal Cancers: Analyses Based on SEER Database [J/OL]. Cancers, 2020, 12 (10): 3–7. https://www.mdpi.com/2072-6694/12/10/2802.

[5] Cabrera S, Bebia V, Acosta U, et al. Survival outcomes and prognostic factors of endometrial stromal sarcoma and undifferentiated uterine sarcoma [J]. Clinical and Translational Oncology, 2021, 23: 1210–1219.