参赛队号：（由大赛组委会办公室填写）

2021年（第七届）全国大学生统计建模大赛

参赛学校:	中南大学
论文题目:	基于半监督学习的脑干海绵状血管瘤手术效果研究
参赛队员:	王佳宁　刘桢杰　王枚
指导老师:	徐宇锋

目　录

表格清单

插图清单

摘　要

大数据时代下，基于临床医疗数据的建模可作为辅助手段，为制定医疗方案提供技术支持。本文以脑干海绵状血管瘤临床数据为研究对象，建立分类模型用于术后远期临床效果预测。全文共分为三个部分。

首先，量化原始数据并确定因变量。根据临床背景归纳离散型变量类型并进行编码，统一连续性变量取值。由术前和随访 KPS 评分变化确定因变量，评分变差则取为 0，否则为 1。最终得到 63 条带标签样本，其中正样本 46 条。

其次，填补缺失值与平衡样本。本文采用 Kmeans-KNN 和随机森林两种方法填补缺失值，比较基于两者的模型 AUC 值可知前者比后者高 0.034。由于正负样本数不一致且存在离散型变量，需用 SMOTE-NC 算法处理得到平衡化样本。

然后，基于半监督学习方式建立分类模型。综合考虑带标签与无标签样本训练模型，对比性能评价指标。加入无标签样本训练后的逻辑回归与 XGBoost 在测试集上的 AUC 值比未加入的 AUC 值分别高 0.015 与 0.056。

本文的研究工作从实际数据出发，建立可靠度较高的术后远期临床效果预测模型，对于该疾病的诊断提供了一定辅助支持。

关键词：脑干海绵状血管瘤；远期临床效果预测；缺失值填补；样本平衡化；Self-Training 半监督学习

Abstract

In the era of big data, clinical medical data model can be used as an auxiliary method to provide technical support for making medical plans. Based on the clinical data of brainstem cavernous hemangioma, a classification model was established to predict the long-term clinical effect after operation.

First, quantify the original data and determine the dependent variable. According to the clinical background, summarizing discrete variables' types , and unifying continuous variables' values. The dependent variable was determined by the change of KPS score before operation and follow-up. If the score deteriorated, it was 0, otherwise it was 1. Finally, 63 label samples were obtained, of which 46 were positive samples.

Then, fill in missing values and balance samples. We use Kmeans-KNN and random forest to fill the missing value. The comparison shows that the AUC value of the model based on the former is 0.034 higher than that of the latter. Because the number of positive and negative samples is inconsistent and there are discrete variables, SMOTE-NC algorithm is used to balance 92 samples.

Finally, establish a classification model based on semi-supervised learning. Considering the training models of labeled samples and unlabeled samples, The AUC values of logistic regression and XGBoost after training with unlabeled samples are 0.015 and 0.056 higher than those without training. This work establishes a highly reliable long-term clinical effect prediction model, which provides some auxiliary support for the diagnosis of this disease.

Key words: Cavernous hemangioma of brain stem; Long-term clinical effect prediction; Missing value filling; Sample equalization; Self-Training semi-supervised learning

基于半监督学习的脑干海绵状血管瘤手术效果研究

一、引言

（一）研究背景与意义

海绵状血管瘤是一种低流量血管畸形，发生在脑干的海绵状血管瘤被称为脑干海绵状血管瘤（BSCM）。脑干海绵状血管瘤占所有颅内海绵状血管瘤的 4%-35%[1]，与发生在颅内其他部位的海绵状血管瘤相比较，脑干海绵状血管瘤更容易发生反复出血现象。目前,关于脑干海绵状血管瘤的治疗方案主要有伽玛刀治疗、保守治疗以及显微手术治疗等，其中，显微手术是对患者病情改善率最高的治疗方法。由于脑干海绵状血管瘤所处部位深，与重要神经结构毗邻[1]，因此手术时存在风险高、难度大和术后患者恢复状态难以预测等问题。故利用统计学模型找出与手术紧密相关的指标并帮助医生预判患者术后恢复状况是具有重要意义的研究课题。

近年来，利用统计学方法对脑干海绵状血管瘤进行的研究越来越多，目前的研究主要集中在利用传统的统计学方法如卡方检验、t 检验等来判断某些指标是否对脑干海绵状血管瘤具有影响，但由于显著性水平设置的个体差异性常会使得这些检验结果具有很大的主观性。随着机器学习的发展，利用机器学习的方法探索隐藏在数据中的信息，找出与脑干海绵状血管瘤相关的指标，并基于这些指标建立相关的机器学习模型，不仅能为患者制定个性化的手术策略，而且还能够促成有关脑干海绵状血管瘤手术治疗的详细指标诊断库的形成。也就是说，将机器学习方法引入脑干海绵状血管瘤的手术治疗中，不仅能消除医生在为患者制定手术策略上的盲目性，提高患者手术成功率，同时可以制定对于完善脑干海绵状血管瘤规范化治疗方案，进行合理的针对性护理干预，对于提高患者术后的生活质量有着非常重要的临床意义。

在实际的医学研究中，由于患者隐私或随访走丢等，常会使得收集的数据集

存在大量缺失值，若将有缺失的样本直接删除，则会造成样本信息的大量丢失，而且样本分布不均衡，存在大量无标签样本也是该类数据集中广泛存在的问题。脑干海绵状血管瘤的手术难度较高，患者数量较少，数据搜集难度较大，因此数据具有非常大的临床价值和统计学价值，但是保留下来的数据存在许多问题，如有大量缺失值、数据不平衡和标签缺失等。如何对医学数据集进行预处理，量化数据，构造因变量，合理的填补缺失值，平衡数据集，并利用大量无标签样本所携带的信息构建分类模型，尽可能实现数据价值的最大化利用，在统计学和医学上无疑都是一个有价值的研究课题。

（二）国内外研究现状

1. 脑干海绵状血管瘤研究现状

脑干海绵状血管瘤（BSCM）是病灶位于脑干的海绵状血管瘤（CM）的一种，与其他海绵状血管瘤相比，脑干海绵状血管瘤更容易出现出血症状。目前脑干海绵状血管瘤的治疗方案主要有伽玛刀治疗、保守治疗以及显微手术治疗等。

2011 年孙季冬等[2]回顾性分析 2008 年 9 月到 2010 年 9 月的 27 例 BSCM 患者的临床资料和预后，发现术后随访的 25 例患者中，伽玛刀组、保守组和显微手术组的病情改善率分别为 4/6、6/7、9/12，保守治疗组的改善率最高，但作者认为这很可能与保守治疗的病人病情较轻有关。2015 年陈见清等[3][3]指出，BSCM 患者是否能够采取手术方法进行治疗需要术前进行风险评估，但当符合手术指征时，相较于伽玛刀治疗和保守治疗，手术治疗更能改善远期预后。2016 年宋国智等[4]利用统计学中的配对 t 检验、卡方检验以及 Cox 模型探讨 Kawase’s 入路切除脑干海绵状血管瘤的临床疗效，发现该入路对于肿瘤位于中脑下部腹外侧、桥脑腹外侧和侧方的海绵状血管瘤效果显著且手术入路、术前是否再出血、肿瘤大小、病变部位和术中切除肿瘤程度对脑干海绵状血管瘤有重要影响。2018 年张力等[5]利用重复测量方差分析和卡方检验对南京军区南京总医院神经外科 2008

年到 2018 年收治的 25 例显微手术治疗患者进行术前术后 KPS 评分，研究表明显微手术是治疗 BSCM 的有效方法，能够改善患者的功能状况。

2. 半监督学习研究现状

半监督学习是近十年来发展迅速的一种介于监督学习和无监督学习之间的新型机器学习方法。其基本思想是在数据集同时存在带标签和无标签样本时，训练模型通过对无标签样本加以利用进行学习从而提高模型性能。

总体来看，对半监督学习的研究主要可以分为三个阶段[6]。第一阶段是半监督学习的起源阶段。对半监督学习的研究最早可追溯到 20 世纪 90 年代，由于相较于带标签样本，无标签样本的获取通常更容易、成本更低，于是开始有学者尝试在训练模型时利用无标签样本来提高模型性能，其中比较著名的为半监督支持向量机和协同训练。1998 年 Vapnik[7]在统计学习理论一书中提出半监督情形下的支持向量机算法：半监督支持向量机。在带标签样本数量较少的情况下，支持向量机的决策边界不应该穿过样本密度较高的区域，因此半监督支持向量机通过在目标函数上加入无标签样本错分代价来规范、调整决策边界。协同训练是 Blum[8]提出的一种多视角学习算法，该算法假设数据集具有两个充分冗余的视图，在两个视图上分别利用带标签样本训练出一个分类器，然后再利用这两个分类器对从无标签样本中随机抽选的若干样本进行标记并把自己认为标记置信度较高的无标签样本连同其对应的分类标记添加到另一个分类器的数据集中，最后所有的分类器利用更新后的数据集进行二次训练，协同训练过程不断迭代进行直至所有的无标签样本都进行了标记或达到停止条件。

第二阶段是半监督学习的成熟阶段。由于上述两种方法都存在各自缺陷，于是人们开始尝试利用其他方法对无标签样本进行学习，直到 2000 年后，半监督学习才正式发展为有别于监督学习和无监督学习的一类独立的机器学习方法，这个时期影响力较大的算法包括混合模型、自训练以及图论半监督等。2000 年

Nigam 等[9]将 EM 算法和朴素贝叶斯分类器混合对带无标签的文本进行分类，实验结果表明，当数据满足模型假设时，可以取得较好的分类效果。2005 年 Rosenberg 等[10]提出一种基于自训练的半监督目标检测系统，证明了自训练模型可以获得与使用更大的带标签数据集以监督学习方法训练同等的结果。图论半监督算法将所有样本以及其对应的关系表示为一个无向图，其中图的结点为数据样本点，图的边表示两两样本之间的关系，基于图的半监督学习的优化目标就是要保证在已有标签样本上的结果尽量符合且满足流型假设。

第三阶段是半监督深度学习的发展阶段。半监督深度学习是半监督学习和深度学习结合的产物，近年来，随着深度学习的发展，其在图像识别、文本分类、自然语言处理等领域均取得优异成果，但由于训练深度学习算法所需要的样本量往往比传统的机器学习大得多，因此人们开始将半监督学习引入深度学习中。生成式对抗网络是目前使用最为广泛的半监督深度学习算法，它主要通过让生成器和判决器相互竞争达到平衡状态来无监督的训练网络。

二、理论基础

（一）缺失值填补算法

数据缺失在现实数据中非常常见。目前，对于缺失值的处理方法基本上分为四类：删除、权重法、填补和不处理[11]。本次研究的数据为临床数据，搜集难度大，时间跨度长，为了尽可能保留数据的价值，选择对缺失值进行填补。缺失值的填补分为简单处理填补和复杂处理填补，常用的简单处理填补方法有平均数填补和中位数填补，复杂处理填补有 K-近邻填补、随机森林填补、聚类填补和多重插补法填补。本次研究的临床数据是包括定性数据和连续数据的混合型数据，且部分特征缺失值较多，因此选择使用复杂处理填补。在无法得知哪种复杂处理填补的效果最好的前提下，本文考虑使用 K-近邻填补和聚类填补相结合的填补

方法和随机森林填补，并对两种方法填充后数据的模型预测效果进行对比。

1. 随机森林算法

随机森林是 2001 年由 Breiman 首次提出一种集成学习算法[12]。其本身并不是一个单独的机器学习算法，而是通过构建多个模型，集成所有模型的建模结果。随机森林算法的具体步骤主要分为如下四步，

Step1：用 bootstrap 方法生成 m 个训练集，分别训练 m 棵决策树。

Step2：在每一棵决策树的结点选择特征进行分裂时，在特征中随机抽取一部分特征，在抽取的特征中进行选择。

Step3：每一棵决策树都尽最大程度生长，并且没有剪枝过程。

Step4：将生成的多棵决策树进行决策。对于分类问题，由多棵决策树投票决定分类结果；对于回归问题，由多棵决策树的预测值均值决定回归结果。

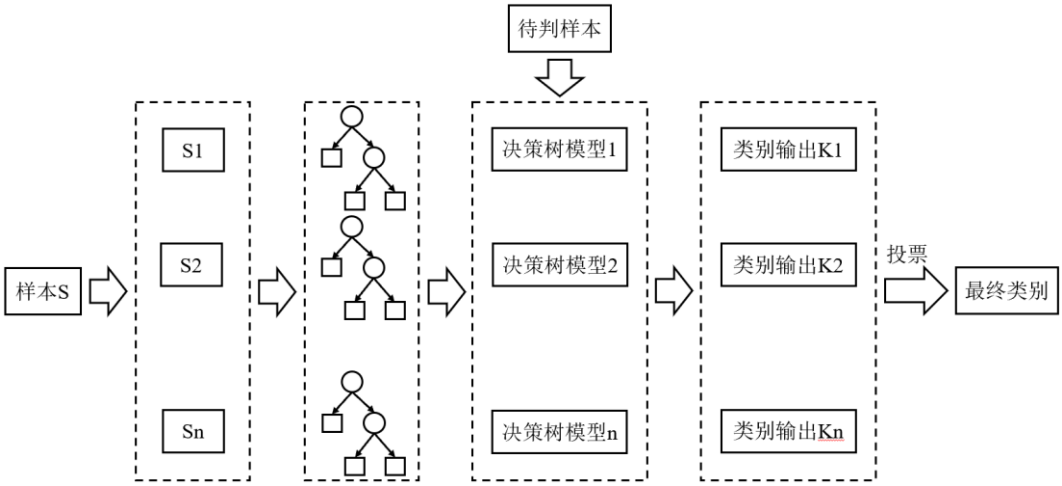


图 1　随机森林算法原理

2. MissForest 算法

MissForest 是一种利用随机森林来填补缺失值的非参数方法，由 Stekhoven 和 Buhlmann 于 2012 年提出。其具体思路和步骤有以下几步，算法的伪代码如表 1 所示[13]。

（1）假设数据$X = (X_1, X_2, \ldots\ldots, X_p)$为$n * p$的一个矩阵，$X_s$为任一含有缺失值的变量。

（2）将数据分成 4 部分：用$y_{obs}^{(s)}$表示X_s的观测值；用$y_{mis}^{(s)}$表示X_s的缺失值；用$x_{obs}^{(s)}$表示X_s观测值以外的其余观测值；用$x_{mis}^{(s)}$表示X_s的缺失值以外的其余观测值；

（3）使用随机森林训练出$y\sim x$的模型，进行缺失值预测即可。但是不止X_s存在缺失值，其他变量也可能存在缺失值，此时采用迭代的方式来求解。先对缺失值做一个初始的猜测，比如用均值或中位数填充，然后按照变量的缺失率，从小到大排序，先对缺失率小的变量使用随机森林填补其缺失值，最后进行迭代，直到最新的一次填补结果与上一次的填补结果不再变化（或变化很小）时停止。

表 1　MissForest 算法伪代码

算法 1 MissForest 算法
输入：$n*p$的数据矩阵$X=(X_1,X_2,\ldots\ldots,X_p)$和停止条件$\gamma$
输出：填补后的数据矩阵X^{imp}
1：对缺失值做初始猜想
2：$k\leftarrow$ 矩阵X中关于缺失变量缺失率排序的索引变量
3：While not γ do:
4：　$X_{old}^{imp}\leftarrow$ 存储初始猜想的矩阵
5：　for s in k do:
6：　　构建$y_{obs}^{(s)}\sim x_{obs}^{(s)}$ 的随机森林
7：　　用$x_{mis}^{(s)}$预测$y_{mis}^{(s)}$
8：　　$X_{new}^{imp}\leftarrow$ 使用预测后的$y_{mis}^{(s)}$更新填补矩阵
9：　end for
10：　update γ
11：end while
12：return 已经填充的数据矩阵X^{imp}

（二）样本平衡化算法

在数据集 S 中，如果不同类别的样本数量差距很大即$N_i \gg N_j$时，就称 S 为不平衡数据集。目前针对不平衡数据集的研究，大致可以分为算法层面和数据层面两种。算法层面的方法是指对算法进行改进以使其适应不平衡数据集的学习，常用的算法主要有集成学习、单类别学习以及代价敏感学习等。但利用算法层面方法来平衡数据集时常常会面临很多新的问题，如何改进算法才能使得少数类样本被重视等。数据层面的方法主要是利用采样技术对数据集进行重构，以此达到

改变样本数量分布的目的。常用的采用技术主要有欠采样和过采样。欠采样方法主要有随机欠采样、Tomek links 和 NearMiss 等，过采样方法主要包括随机过采样和 SMOTE 等。由于 SMOTE 算法简单易行且不会损失原始数据集所携带的信息，在很多研究中取得了较好的效果。因此，本文采用将 SMOTE 算法的变体 SMOTE-NC 算法来对数据集进行平衡化。

1. SMOTE 算法

SMOTE 是 Chawla 于 2002 年提出的一种经典的综合少数类过采样算法[14]。该算法的主要目的是通过线性插值的方式增加少数类样本的数目从而使得数据集达到平衡。其原理如图 2 所示[15]，下面给出 SMOTE 算法的具体步骤：

Step1：针对不平衡数据集 S，将其划分为多数类S_{maj}和少数类S_{min}；

Step2：对于每一个少数类样本，计算其 K 近邻；

Step3：根据数据集的不平衡比例决定每个少数类样本待合成的新样本个数 N。然后对每一个少数类样本，从其 K 近邻中随机选取 N 个近邻。假设在利用样本x合成新样本时选择的近邻为x_n，再按照公式$x_{new} = x + rand(0,1) * (x_n - x)$ 构建新的样本。

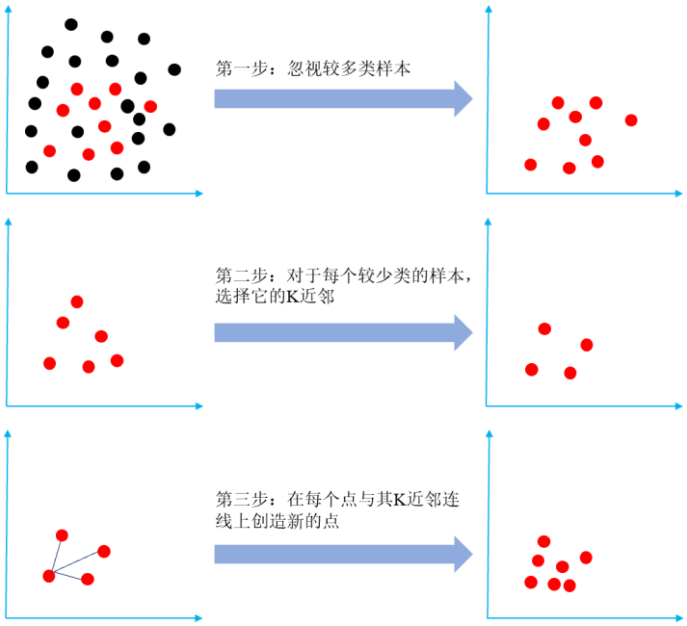


图 2　SMOTE 算法原理

2. SMOTE-NC 算法

从上一节对 SMOTE 算法的介绍可以发现，由于 SMOTE 算法在计算 K 近邻时采用的距离计算方法为欧氏距离，所以当数据集存在离散变量时，传统的 SMOTE 算法不太适用。为使得算法能应对数据集同时存在连续和离散变量的情形，Chawla 在提出 SMOTE 算法的同时给出了该问题的解决算法——SMOTE-NC 算法[14]。

表 2　SMOTE-NC 算法伪代码

算法 2 SMOTE-NC 算法
输入：数据集$S = S_{min} + S_{maj}$,近邻个数 K，采样倍率 N
输出：平衡后的数据集S'
初始化：新样本数据集$S_{new} = \emptyset$
1：将原始数据集 S 划分为少数类样本集S_{min}和多数类样本集S_{maj}
2：for x_i in S_{min}：
3：　　n=1
4:　　从S_{min}中计算x_i的 K 近邻X_k。在计算两两样本之间的欧氏距离时，离散变量的差异用所有连续变量的标准差的中值表示
5：　　While $n \leq N$：
6:　　从X_k中随机选择一个样本，记为x_n
7：　　合成新的样本x_{new}。其中新样本的连续变量取值用$x_i + rand(0,1) * (x_n - x_i)$生成，离散变量取值用 K 近邻的众数填充
8:　　$S_{new} = S_{new} \cup x_{new}$
9:　　n=n+1
10:　end for
11: $S' = S \cup S_{new}$

SMOTE-NC 算法和 SMOTE 算法的基本思想大体一致，主要在以下方面进行了改进，其算法伪代码如表 2 所示。

（1）中值计算。计算少数类样本的所有连续变量的标准差中值，如果一个样本与其潜在的最近邻之间的离散变量取值不同时，将该中值作为惩罚加入这两个样本欧氏距离的计算中。

（2）最近邻计算。在计算每个少数类样本的 K 近邻时，对于每一个取值不同的离散特征，计算欧氏距离时将包含上述计算的标准差的中值。

（3）合成新样本。合成的新样本的连续变量的取值与 SMOTE 算法描述一

致，离散变量取值由其 K 个近邻的众数填充。

如给定两个样本F_1=[1,2,3,A,B,C]、F_2=[4,5,6,A,D,E],则F_1与F_2之间的欧氏距离为：$\sqrt{(4-1)^2+(5-2)^2+(6-3)^2+Med^2+Med^2}$，其中 Med 为该少数类样本连续变量的标准差中值。

（三）半监督学习框架

利用机器学习模型解决实际问题时，传统的做法是只考虑在带标签的样本上评价模型的效果。但如果带标签的样本量较少，此时的模型会存在过拟合现象，导致泛化性能不高。如果仅使用无标签数据集进行无监督学习，又忽略了带标签数据的重要价值。此外通常情况下都是无标签数据比较多，对无标签数据进行人工标注是一件耗时耗力的任务。因此十分有必要考虑半监督学习将少量带标签数据与大量无标签数据结合起来提升分类器性能，弥补无监督学习与监督学习各自的缺点。半监督学习除了具有完成小的训练集下的分类，以减少训练时间或代价的优点外，还能利用无标签测试数据的信息调整分类器参数，从而提高分类器的自适应性。

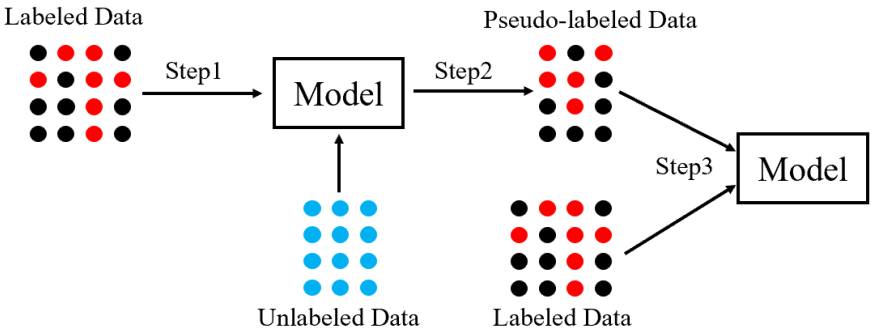


图 3　半监督学习算法流程图

半监督学习算法根据问题不同主要分为分类和回归两大类。半监督分类算法的流程如图 3 所示，主要过程分为三个部分。第一步，利用带标签数据训练好模型；第二步，利用训练好的模型预测无标签数据的标签；第三步，综合无标签数据与带标签数据继续调整模型参数至最优结果，本文采用的方法属于半监督分类算法中的 Self-Training 算法，主要以半监督的方式训练模型。把分类

器对无标签数据的预测作为无标签数据的伪标签(Pseudo Label)，再结合伪标签进行训练。具体算法描述如表 3 所示。

表 3　Self-Training 半监督学习算法

算法 3 Self-Training 半监督学习算法
输入：数据集$S=(N_L,T,H)$，类别数C，每个类别样本数K
输出：分类模型Ω
1：将原始数据集 S 划分为有标签样本集$S_L=(N_L,T,H)$和无标签样本集$S_U=(N_U,T,H)$
2：在有标签样本集$S_L=(N_L,T,H)$上训练模型，调整参数以取得较优结果
3：利用训练好的模型对所有无标签样本X_i进行预测，得到类别概率P_i与伪标签PL_i
4：$For\ X_i\ in\ S_U$：
5：　　$if\ P_i \geq threshold$：
6：　　$S_L=S_L+(X_i,PL_i)$　//将类别概率 P 大于阈值的样本X_i与伪标签PL_i加入S_L中
7：　　$S_U=S_U-X_i$　　//将类别概率 P 大于阈值的样本X_i从S_U中剔除
8：　　$else$：
9：　　$break$
10：　在新数据集S_L上重新训练模型，并对S_U中所有样本X_i进行预测，得到类别概率P_i与伪标签PL_i
11：E$nd\ for$
12: 输出分类模型Ω，算法停止

三、基于半监督学习的 BSCM 远期临床效果预测

（一）数据介绍与因变量定义

1. 数据来源

本文所研究的数据来源于 2008 年至 2020 年在中南大学湘雅医院神经外科就诊并进行显微手术治疗的共 98 例脑干海绵状血管瘤患者的临床数据资料。其中的数据资料包括患者的影像学资料、手术相关特征、几个重要日期、神经功能变化、肌力变化和患者 KPS 评分变化等。患者纳入标准如下：

（1）患者符合脑干海绵状血管瘤的诊断标准，CT/CTA，MR 等影像学资料诊断明确，并在术后病理学检查中报告为海绵状血管瘤；

（2）患者均符合手术指针且在中南大学湘雅医院进行 BSCM 显微手术治疗；

（3）患者临床资料均完善。

由于近十年来，该院的脑干海绵状血管瘤的治疗方案已经非常成熟，且术者

之间的技术差异不大，对于术后恢复的影响可以忽略不计，因此认为这 98 例患者具有统一的纳入标准，即不会因为时间变化，医疗设备和条件导致患者样本间存在系统性差异。

2. 数据量化

根据研究的目的，纳入考量的变量包括患者性别、年龄、肿瘤部位、肿瘤大小、手术入路、是否出血、是否气切、是否再出血、是否脑积水、首次症状距离手术天数、ICU 住院天数、手术后住院天数、颅神经事件、长期肌力变化、短期肌力变化、术后 72 小时 KPS 评分、长期 KPS 评分变化、术前颅神经症状和术后颅神经症状共 19 个变量。

对于连续变量，不需要进行量化处理；对于文本变量，可以对其进行词云图分析，下文将进行详细描述；对于定性变量的量化处理较为复杂，需要根据变量特征一一进行量化。

患者性别分为两类，女性刻画为 0，男性刻画为 1。肿瘤部位分为中脑、中脑及桥脑、桥脑、桥脑及延髓和延髓共五类，依次刻画为 1、2、3、4 和 5。手术入路分为幕下小脑上、远外侧、后正中、枕下乙状窦后、翼点和纵裂入径共 6 类，依次刻画为 1、2、3、4、5 和 6。是否出血、是否气切、是否再出血和是否脑积水都分为两类，否刻画为 0，是刻画为 1。颅神经事件分为无变化、术后有术前无、术前有术后不变、术前无术后有和术前有术后新发症状共五类，依次刻画为 0，1，2，3 和 4。长期肌力变化和短期肌力变化都分为好转、不变和变差三类，分别刻画为 1，0，-1。长期 KPS 评分变化分为未好转和好转两类，分别刻画为 0 和 1。

3. 数据结构分析

本次研究的数据包括 98 个样本和 19 个变量。其中包括 6 个连续变量、11 个定性变量和 2 个文本变量。从对量化后的数据进行初步统计分析来看，不存在异

常值的问题，但是存在大量缺失值。长期肌力变化、长期 KPS 评分变化、肿瘤大小、首次症状距离手术天数、术后 72 小时 KPS 评分、肿瘤部位、手术入路、颅神经事件和短期肌力变化共 9 个变量存在缺失，其中长期肌力变化、长期 KPS 评分变化和肿瘤大小存在缺失值较多，分别为 38 个、35 个和 30 个。因此在本次研究中，缺失值的处理至关重要，决定了数据质量以及模型预测结果的好坏。

4. 因变量的确定

本文研究的目标是脑干海绵状血管瘤显微手术远期临床效果，手术的远期效果对应于患者随访时的 KPS 评分，即 Karnofsky 功能状态评分标准。KPS 评分刻画了患者的生活自理能力，即患者的生活质量，常认为患者的 KPS 评分越高，患者的生活的自理能力越强，生活质量越高。一般若患者的 KPS 评分大于 80，认为患者术后状态较好，存活时间较长。

表 4　KPS 评分对照表

KPS 评分	生活自理能力
80 分以上	非依赖级，生活可自理
50～70 分	半依赖级，生活半自理
50 分以下	依赖级，生活不能自理

如何判断患者术后的状态是否好转，要对比随访的 KPS 评分较术前的 KPS 评分是否有提高或者是否保留了术前较高的生活自理能力，因此本文的因变量确定为患者长期 KPS 评分变化。。

表 5　因变量对照表

长期 KPS 评分变化	患者术后状态	因变量标签
ΔKPS>0	好转	1
ΔKPS=0,随访 KPS≥80	好转	1
ΔKPS=0,随访 KPS<80	未好转	0
ΔKPS<0	未好转	0

对于长期 KPS 评分变化来说，存在三种情况，评分提高、评分不变和评分下降。评分提高，则认为患者术后恢复较好，状态好转；评分下降，则认为患者术后恢复较差；评分不变存在两种情况，一是患者评分虽然没有改变，但是随访

的 KPS 评分依然大于 80，则认为该患者术后恢复的不错，状态好转，二是随访的 KPS 评分小于 80，则认为该患者术后恢复不太好，状态未好转。

长期 KPS 评分变化为二分类定性变量，包括患者的生活自理能力好转，即患者术后的状态好转，刻画为 1；患者的生活自理能力未好转，即患者的术后状态未好转，刻画为 0。本文研究的因变量是 0-1 二分类定性变量，研究的问题其本质是一个二分类定性变量的预测问题。本次研究中涉及的所有变量到目前已经完全介绍清楚，如表 6 所示。

表 6　变量说明表

变量类型		变量名	详细说明	取值范围
因变量		是否好转	定性变量(2 个水平)	好转/好转
自变量	患者信息	性别	定性变量(2 个水平)	女性患者/男性患者
		手术入路	定性变量(6 个水平)	幕下小脑上/远外侧/后正中/枕下乙状窦后/翼点/纵裂入径
		年龄	单位：岁	4-82
		首次症状距离手术日期	单位：天	1-1825
		ICU 住院天数	单位：天	0-30
		术后住院天数	单位：天	2-211
	肿瘤特征	部位	定性变量(5 个水平)	中脑/中脑和桥脑的中间地带/桥脑/桥脑和延髓的中间地带/延髓
		大小	单位：立方毫米	240-31500
	相关症状	是否出血	定性变量(2 个水平)	出血/未出血
		是否脑积水	定性变量(2 个水平)	脑积水/未脑积水
		是否气切	定性变量(2 个水平)	气切/未气切
		是否再出血	定性变量(2 个水平)	再出血/未再出血
	恢复状况	颅神经事件	定性变量(5 个水平)	无明显变化/术前有术后无/术前有术后不变/术前无术后有/术前有术后新发症状
		术后肌力变化	定性变量(3 个水平)	变差/未变/变好
		随访肌力变化	定性变量(3 个水平)	变差/未变/变好
		术后 72hKPS	单位：分	20-90
		术前颅神经	文本变量	例：右上肢无力
		术后颅神经	文本变量	例：双侧瞳孔不大

（二）缺失值填补

1. 缺失数据可视化

上文中提到本次研究的数据中存在大量缺失值，需要对缺失值进行填补。确定长期 KPS 评分变化为因变量，剔除两个文本变量，为了进一步了解变量的缺失情况，对剩下的 17 个变量其进行可视化处理。

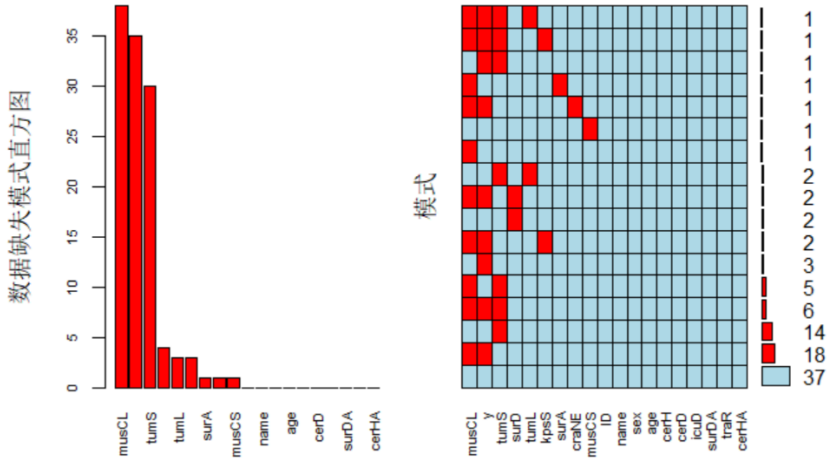


图 4　缺失数据可视化条形图

从图 4 中左图可以看出，包括因变量在内共 9 个变量存在缺失值，依次为长期肌力变化（musCL）、长期 KPS 评分变化（y）、肿瘤大小（tumS）、首次症状距离手术天数（surD）、术后 72 小时 KPS 评分（kpsS）、肿瘤部位（tumL）、手术入路（surA）、颅神经事件（craNE）和短期肌力变化（musCS），且存在缺失值个数依次为 38 个、35 个、30 个、4 个、3 个、3 个，1 个，1 个和 1 个。可以得出，存在较多缺失值的变量为长期肌力变化、长期 KPS 评分变化和肿瘤大小。右图中红色区域显示了缺失变量在样本中的分布情况，如变量长期肌力变化的缺失值分布在第 38～45 个、第 50～55 个、第 56～60 个、第 64～65 个、第 68～69 个、第 92 个、第 94 个、第 95 个、第 97 个和第 98 个样本中。

图 5 只显示了首次症状距离手术天数（surD）和肿瘤大小（tumS）的数据缺失项及这两个数据项之间的交汇图。图中，左边的红色矩形表示缺失肿瘤大小的数据项，但是首次症状距离手术天数数据没有缺失的样本分布，而左边的灰色矩

形显示剩余样本点的分布。类似的，图下方的红色矩形表示缺失首次症状距离手术天数的数据项，但是肿瘤大小数据项没有缺失的样本分布，而灰色矩形同样显示剩余样本点的分布。若缺失值的缺失服从完全随机分布（MCAR），则红色矩形和灰色矩形应该看起来很相似，可以判断首次症状距离手术天数和肿瘤大小的缺失值的缺失服从完全随机分布。

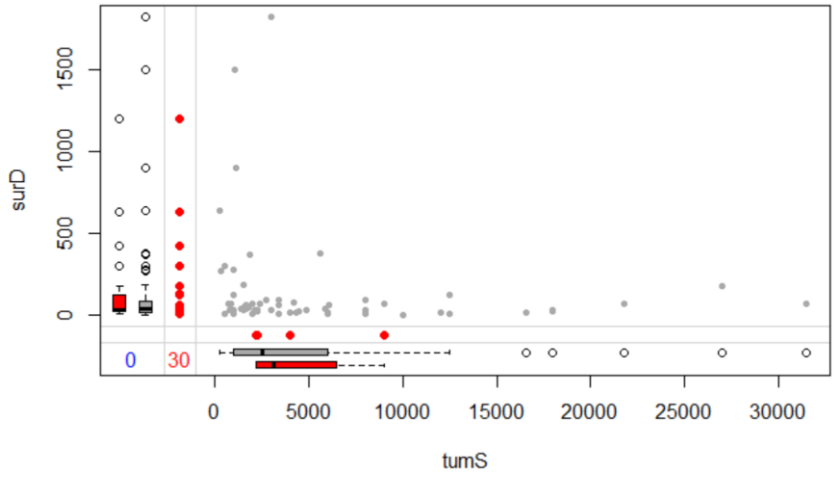


图 5　缺失数据分布统计图

2. 基于 Kmeans-KNN 算法填补缺失值

由上文可知，包括因变量在内共 9 个变量存在缺失值。在缺失值的填补中，不考虑因变量的填补，仅对存在缺失值的 8 个因变量进行填补。由图 4 可知，长期肌力变化和肿瘤大小缺失值较多，且这些缺失自变量中既存在连续变量又存在定性变量，因此考虑 kmeans 和最近邻算法相结合的方法进行填补。

首先剔除长期肌力变化和肿瘤大小缺失值较多的两个自变量，其次剔除只缺失一两个数据的样本，对剩下的样本使用系统聚类确定类别为 4，再使用 kmeans 聚类输出每个样本的类别；然后计算之前剔除的样本与这 4 个类中心的距离，找到距离最近的类，确定之前剔除的样本的类别，对于连续变量的缺失值用这个类的平均值填补，对于定性变量的缺失值用这个类的众数填补；最后基于所有样本用最近邻算法填补缺失值最多的自变量，即长期肌力变化和肿瘤大小。

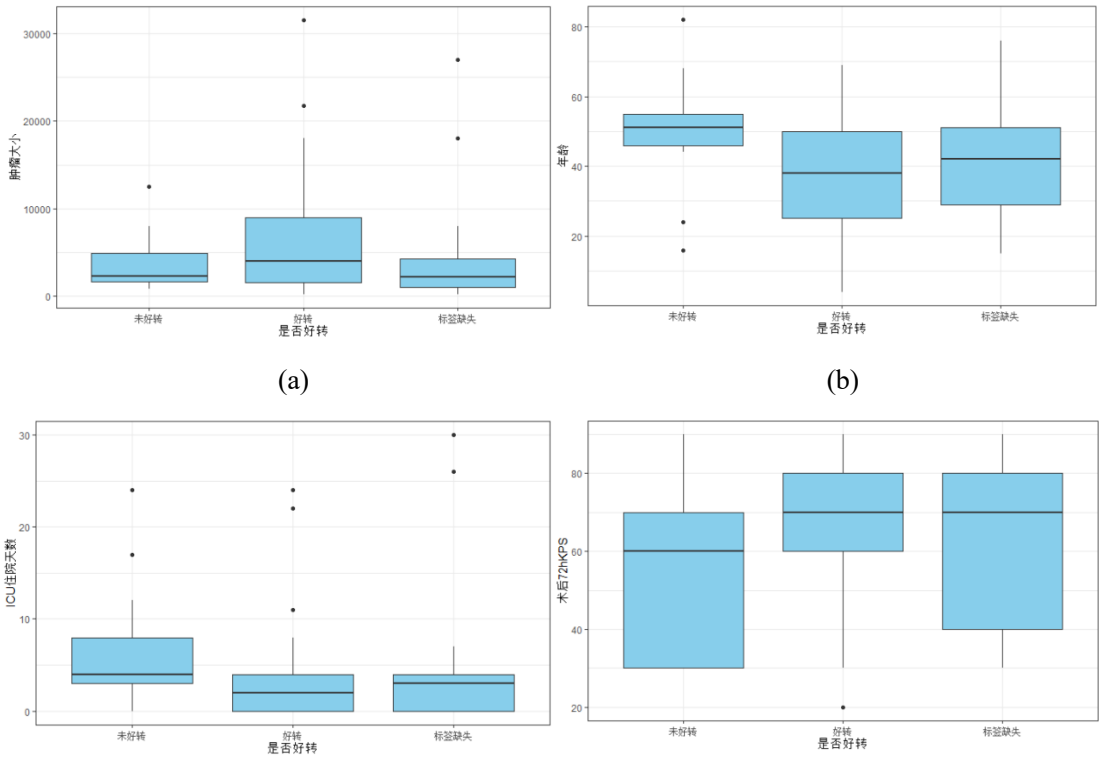
3. 基于随机森林填补缺失值

MissForest 是一种利用随机森林来填补缺失值的方法。本次研究中共有 8 个自变量存在缺失值，具体缺失情况如图 4 所示。使用 MissForest 算法进行缺填补，首先对缺失自变量做一个猜测，连续变量使用均值，定性变量使用众数；然后按照变量的缺失率，从小到大依次进行填补，本次研究中填补的顺序为短期肌力变化、颅神经事件、手术入路、肿瘤部位、术后 72 小时 KPS 评分、首次症状距离手术天数、肿瘤大小和长期肌力变化；最后对于每一次填补过程，一直进行迭代，直到最新的一次填补结果与上一次的结果不再变化（或变化很小）时停止。

（三）描述性统计

1. 基础描述

在完成所有缺失值的填补后，对数据进行描述性统计分析，探究因变量和自变量之间的相关关系，对数据进行初步探索。



(a) (b)

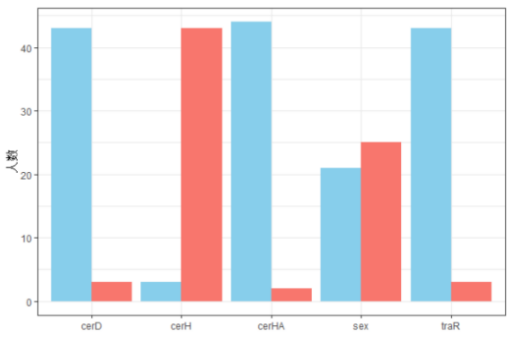
(c) (d)

图 6　术后患者分布箱线图

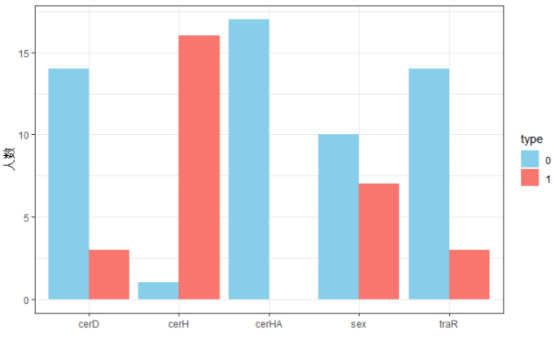
从图 6 三个人群在各个变量上的分布可以看出，经显微手术治疗后，病情好

转与未好转的患者在肿瘤大小、ICU 住院天数上的均值差异较小，在年龄和术后 72hkps 评分上的均值较大。因此可以推测，相较于肿瘤大小和 ICU 住院天数，年龄和术后 72hkps 对患者病情是否好转的影响更大。

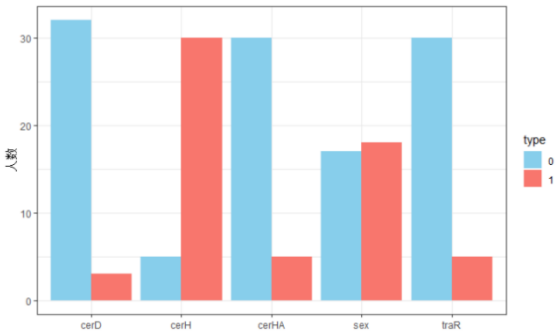
在本文数据集中，肿瘤大小最小值为 $3500mm^3$，最大值为 $31500mm^3$，图(a)表明，好转病人与未好转病人的肿瘤大小均值均在 $5000mm^3$以下，好转病人的四分位距比未好转病人要大，由此可以推断肿瘤大小对术后病情是否好转影响较小。这可能是由于 BSCM 患者大多是因肿瘤破裂急性出血引起的肿瘤体积增大，若出血后能够及时送医就诊，则对病情的影响不大。图(b)中好转患者的平均年龄在 40 岁以下，要比未好转患者的平均年龄小很多，分析其原因不难发现高龄患者的心肺功能通常较差，对手术容忍性较低，术后容易出现各种并发症。图(c)表明约 80%的好转患者 ICU 住院天数少于未好转患者的均值，当患者病情过于严重时，即使通过手术也很难改善其病情。图(d)表明，好转患者的术后 72hkps 评分均值高于未好转患者且四分位距更小，结合表 4 可知术后 72hkps 得分越高就越说明术后患者状态越好，所以术后 72kps 与患者病情是否好转相关性较大。



(a)



(b)



(c)

特征	取值	好转	未好转
cerD	0	93%	82%
	1	7%	18%
cerH	0	7%	6%
	1	93%	94%
cerHA	0	98%	100%
	1	4%	0%
sex	0	46%	59%
	1	54%	41%
traR	0	93%	82%
	1	7%	18%

(d)

图 7　术后患者分布条形图

图 7 分别表示手术后情况好转、未好转以及缺失样本在性别(sex)、是否出血(cerH)、是否脑积水(cerD)、是否气切(traR)和是否再出血(cerHA)五个变量上的分布情况。可以看出，在上述五个变量中，三个人群在性别和是否脑积水、是否气切上的分布差异较大，说明这三个变量与术后病情是否好转存在较大相关关系。其中性别与术后病情相关性最大，这可能是与性别差异导致的激素水平不同有关。

2. 词云图描述

术前颅神经症状和术后颅神经症状是文本变量，其中反馈了大量脑干海绵状血管瘤患者发病时产生的相关症状。词云图是对文本数据中出现频率较高的关键词给予视觉上的冲突，形成关键词的渲染。使用词云图对术前颅神经症状和术后颅神经症状进行描述，能够筛选出脑干海绵状血管瘤患者发病时出现最多的颅神经症状，比较术前和术后的颅神经症状差异，有助于疾病的诊断和治疗，具有实际的临床意义。

(a) (b)

图 8　颅神经症状词云图

从图(a)中可以看出，在患者的术前颅神经症状中，“面部”、“麻木”、“头晕”、“视物模糊”、“反射”和“肢体”等词语出现的频率较高，说明患者的术前的颅神经症状多为面部感觉异常，视物不清，头晕和肢体麻木，对于脑干海绵状血管瘤患者的术前诊断，出现这几种症状时，应该给予更多重视。图(b)中，“眼球”、“左眼”、“右眼”、“瞳孔”和“外展”等，说明患者术后的颅神经症状主要集中在眼部，多发症状几乎都是眼部症状，在术后应该给予患者的眼部功能恢复更多重视，制定个性化的预后措施，进一步提高患者的恢复水平，改善患者的术后生活质量。

（四）样本平衡化

在对数据集进行缺失值填补之后，原始数据集共 63 例带标签样本，其中术后有好转者 46 例、未好转者 17 例，数据不平衡比例大于 2。

表 7 原始数据分布表

类别	样本量	标签	比例
未好转	17	0	0.27
好转	46	1	0.73
总计	63	--	1

对于不平衡数据集，如果直接对其进行分类，分类模型将朝着样本量大的一侧倾斜，使得分类结果不可靠或没有研究价值，故需要先对其进行平衡化处理。由于原始数据集同时存在连续变量和离散变量，因此，如 2.2 节所示，本文将采用 SMOTE-NC 算法对数据集进行平衡化处理。

在利用 SMOTE-NC 算法进行平衡化处理时，由于不平衡比例约为 3，因此将近邻个数设置为 5，最终少数类与多数类的样本比例设置为 1：1。平衡后的样本信息如表 8 所示。

表 8　平衡后数据分布信息表

数据集	类别	样本量	标签	比例
Kmeans-KNN 填补	未好转	46	0	0.5
	好转	46	1	0.5
	总计	92	--	1
MissForest 填补	未好转	46	0	0.5
	好转	46	1	0.5
	总计	92	--	1

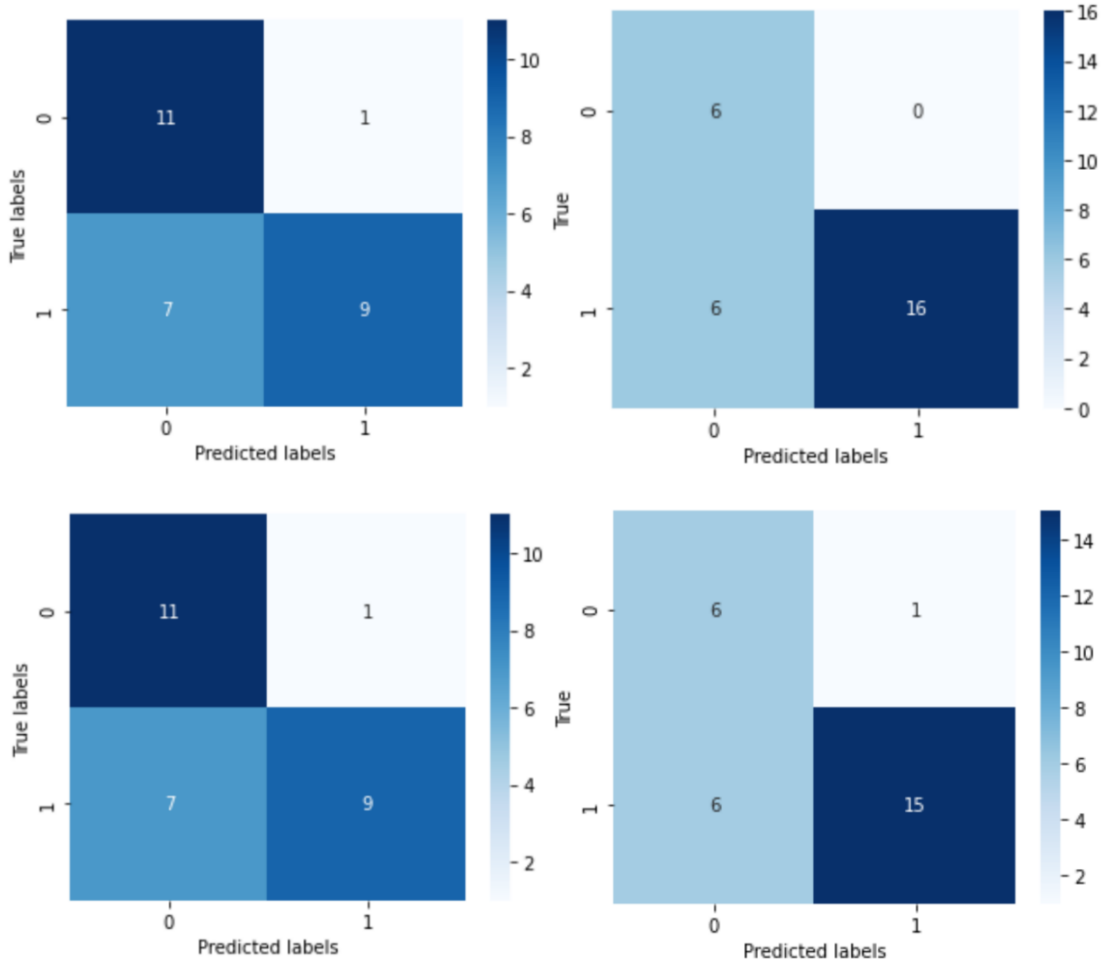
（五）基于 Self-Training 的半监督学习建模

由于脑干海绵状血管瘤是一种非常小众的疾病，因此样本量本身就比较小。为避免只考虑带标签样本建模而忽视无标签样本的珍贵价值，本文将综合使用 98 条原始数据中 63 条带标签样本和 35 条无标签样本，采用半监督学习方法建模对脑干海绵状血管瘤显微手术远期临床效果进行预测，最终提高模型泛化性能。

1. Kmeans-KNN 与随机森林填补缺失值的建模效果比较

在两种处理缺失数据的基础上，分别建立带 L1 正则项的逻辑回归与 XGBoost 模型来比较这两种方法的效果。利用 SMOTE-NC 算法对两个数据集平衡化后，设置训练集与测试集大小比例均为 7:3，再将数据代入模型中得到混淆矩阵与 AUC 值，如图 9 所示。图 9 中左边三幅图是带 L1 正则项的逻辑回归的建模效果，右边三幅图基于 XGBoost 的建模效果。混淆矩阵中上两幅图基于 Kmeans-KNN 填补，下两幅基于随机森林填补。

在图 9 中，L1 逻辑回归下两种处理方法的混淆矩阵是一致的，但是基于 Kmeans-KNN 填补的 AUC 值比基于随机森林填补的 AUC 分别大 0.034 和 0.01。这可能是基于 Kmeans 与最近邻填补的模型在分类时能以更大的概率来判断测试样本的所属类别。因此后续半监督学习将基于 Kmeans 与最近邻填补的模型展开。



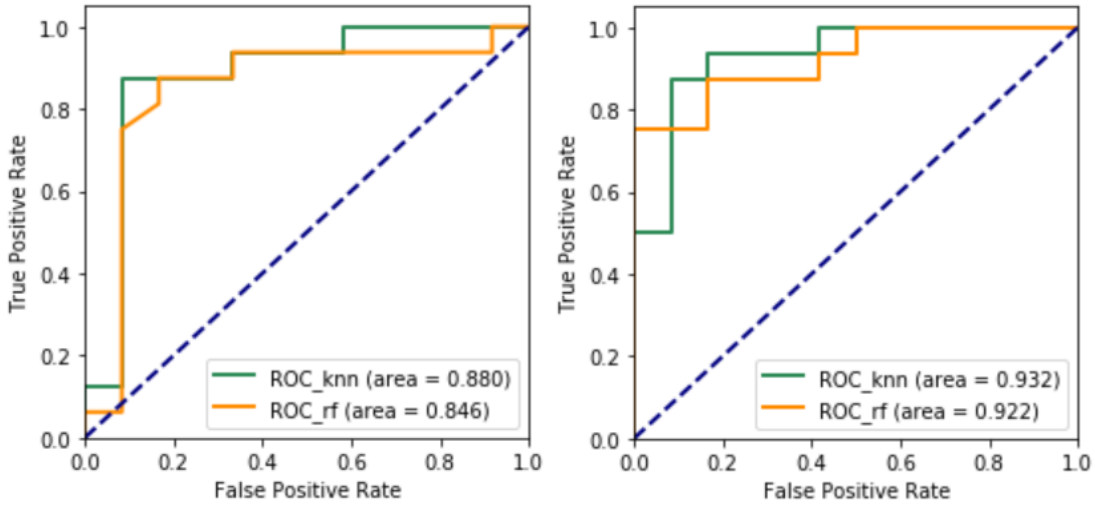


图 9　两种缺失数据的建模效果对比图

2. 基于 Self-Training 的半监督学习建模

基于本部分第一节内容，本文基于 Self-Training 的半监督学习建模过程主要分为如下步骤：

首先，对带标签样本集Ω中所有样本采用 SMOTE-NC 算法平衡化，然后建立带 L1 正则项的逻辑回归和 XGBoost 分类模型，并调节参数以取得较优结果；

其次，利用上述建立好的模型预测无标签样本集Ω_1中所有样本的伪标签，并输出对应的预测概率。对于预测概率大于 0.73 的无标签样本X_i，将其加入到带标签样本集Ω中，其中 0.73 是带标签样本集Ω中因变量为好转的样本数量占所有 63 条样本的比例。

然后，对加入伪标签样本后的样本集重复前两个步骤，直到Ω_1中类别概率没有大于0.73的样本或Ω_1中已无数据后停止循环。对于Ω_1中类别概率没有大于0.73的无标签样本将其全部标记为 0，即作为术后未好转类型。

最后，把所有无标签样本加入Ω中后，再次采用 SMOTE-NC 算法平衡化，建立最终的模型，并比较模型性能指标。

按照上述半监督学习流程完成伪标签后，最终 35 条无标签样本集中 30 条样本的伪标签预测为 1，即为术后好转类型，5 条样本伪标签预测为 0，即为术后未好转类型。将这 35 条样本和对应的伪标签与 63 条带标签的样本集结合起来后

共有 76 条样本标签为 1，22 条样本标签为 0。对总共 98 条样本采用 SMOTE-NC 算法平衡化后，建立最终的模型，输出混淆矩阵与 ROC 曲线图如图 10 所示。

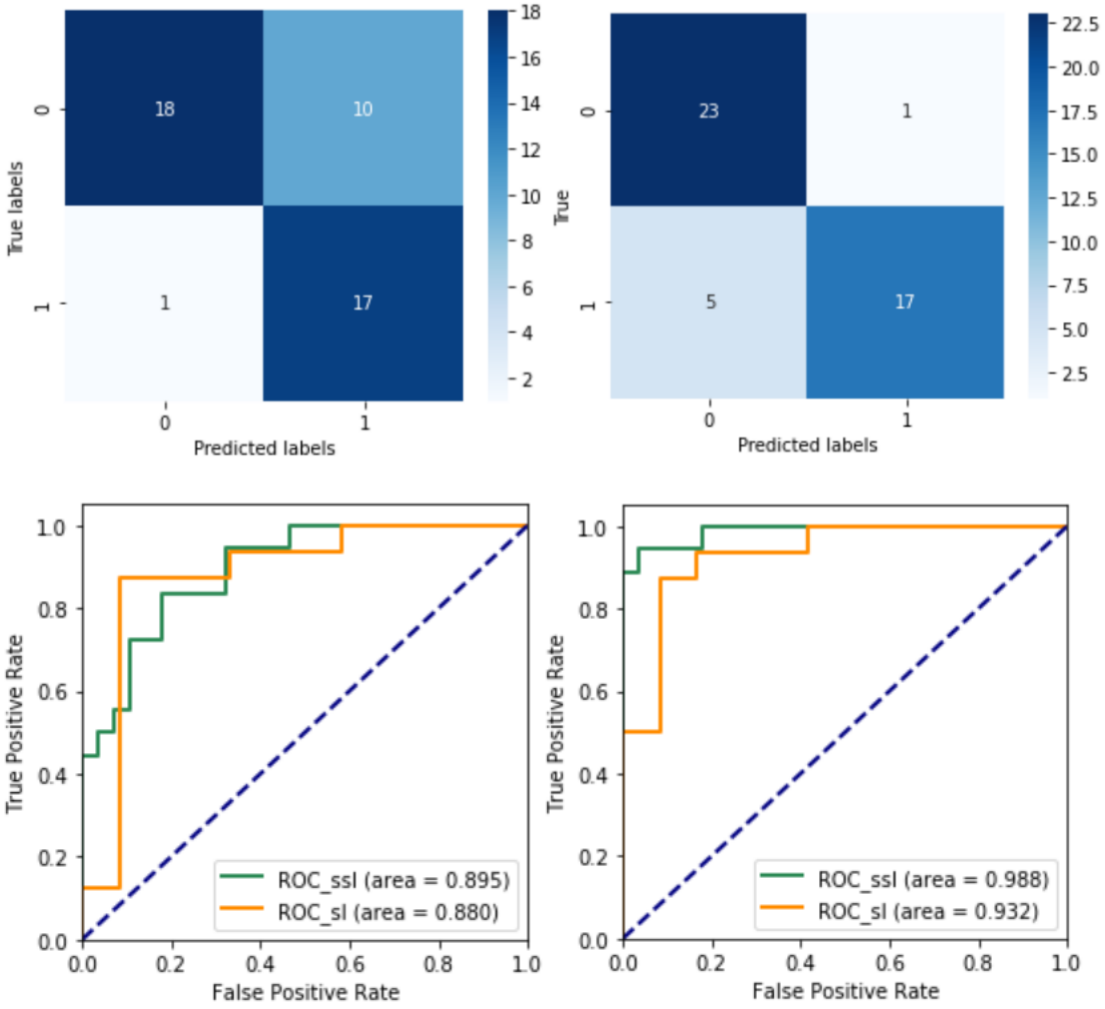


图 10　半监督学习建模效果图

图 10 中上面两幅图是基于带 L1 正则项逻辑回归的建模效果，下面两幅图基于 XGBoost 分类模型的建模效果。由图 10 可知，加入无标签样本训练后的逻辑回归与 XGBoost 在测试集上的 AUC 值分别高达 0.895 与 0.988。相较于图 9，加入无标签样本训练后的逻辑回归与 XGBoost 在测试集上的 AUC 值比未加入无标签样本的 AUC 值分别高 0.015 与 0.056。这就证明通过半监督学习方式将无标签数据集引入到模型训练中能很好的提高模型的性能，为医生在临床诊断过程中提供了一定水平的辅助技术支持。

（六）医学意义分析

为对上述建立的模型进行解释，挖掘其潜在的医学意义，本文接下来将利用 SHAP Value 的方法来对特征的重要性进行排序。

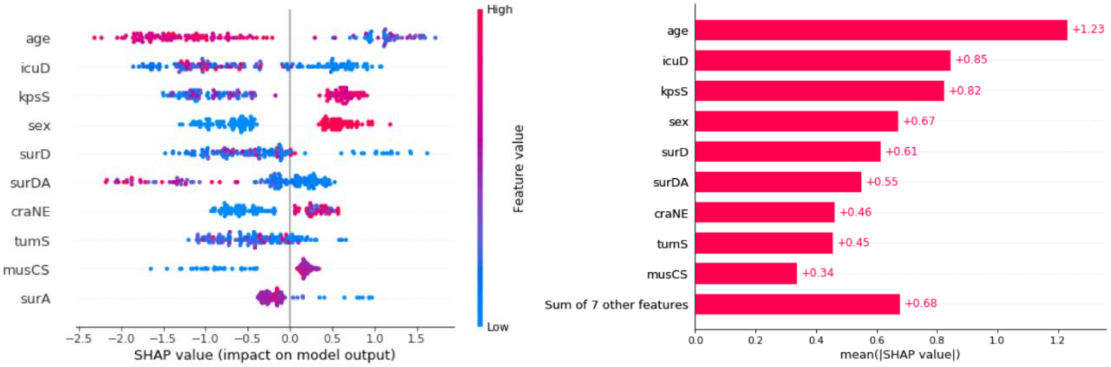


图 11　特征聚合图与特征 SHAP 值

在特征聚合图中，每个点代表一个样本，它在 X 轴上的位置表示特征的 SHAP 值，颜色代表特征的相对大小，红色代表高，蓝色代表低。在特征 SHAP 值中，SHAP 值越高，对应特征越重要。综合可知在用于建模的 16 个特征中，年龄最重要，其次为在 ICU 住院天数和术后 72hKPS 得分。

在实际临床数据建模问题中，往往出现对少数几个变量建立模型即可得到较高的准确率。因此基于特征重要度排序的特征筛选过程对于数学模型与临床实际情况均十分重要。由上述结果可知，在病人进行手术后，医生可以首先根据 SHAP 值较大的几个特征，如年龄、术后 ICU 住院天数与 72hKPS 得分，对患者进行分类。针对不同类别的患者，实施不同的术后恢复治疗方案；针对同一类别的患者，在整体方案一致的前提下，根据其它特征进行微调，做到每一位患者均有属于自己的个性化治疗方案，从而真正在宏观上把控手术效果。这一发现不仅可以降低临床成本，节省医生的精力，还可以对术后效果不好的患者的后续治疗方案进行及时调整，大大节省其治疗时间。

四、总结与展望

（一）总结

本文主要研究工作围绕中南大学湘雅医院神经外科提供的 98 例脑干海绵状血管瘤患者的临床数据资料展开，主要分为如下四个部分：

首先，对原始数据量化并确定因变量。由于临床医学数据的实际情况比较复杂，不适合直接用于建模，需要根据该疾病的治疗背景与临床含义进一步量化。另一方面，因变量并未直接给出，也需要根据实际意义确定。本文通过比较术前 KPS 评分和随访 KPS 评分的变化将因变量确定为好转和未好转两类，来预测该疾病进行手术后的远期临床效果预测。

其次，填补缺失值。量化数据后可以发现因变量和自变量均存在不同程度的缺失值。对于自变量的缺失，本文采用两种处理方法，并在后续中对比建模效果；对于因变量的缺失，本文采用半监督学习方式综合考虑带标签样本与无标签样本，以增加模型的鲁棒性。

然后，采用 SMOTE-NC 算法对样本平衡化处理。利用不平衡数据集建立传统的机器学习模型精度非常不好，最终导致模型的预测结果不可信，因此很有必要在建模之前平衡数据集。本文的自变量中存在许多离散型变量，结合此特点采用 SMOTE-NC 算法进行平衡化处理。

最后，采用半监督学习方式建立预测模型。综合考虑无标签样本与带标签样本以半监督学习方式建立模型，输出模型性能评价指标，并对比未加入无标签样本进行训练的模型性能评价指标，分析结果。最终总结全文，并提出后续方向。

（二）展望

本文针对中南大学湘雅医院神经外科提供的 98 例脑干海绵状血管瘤患者的临床数据资料，改变了以往单纯利用传统的统计学方法对其进行研究的思路，利用机器学习的技术对其进行预处理，量化数据，构造因变量，合理的填补缺失值，

平衡数据集，并利用大量无标签样本所携带的信息构建分类模型，得到较好的分类效果。为后续学者提供了进一步研究思路：

第一，针对数据集的不平衡问题。本文处理的是二分类情况下的不平衡数据集，但在二分类情况下预测效果好的模型不一定在多分类情况下也表现出同样的优越性，因此，后续研究可以针对如何处理多分类情况下的数据集展开。

第二，针对数据的分类预测。本文采用半监督学习方式建立预测模型，但由于 Self-Training 不是一个鲁棒性的学习算法，因此在算法迭代训练的后期，可能仍然会存在无标签样本被标错的风险。因此，后续研究可以考虑从如何减少标错样本或利用无监督学习方式先对无标签样本进行处理等展开。

参考文献

[1] 冯孟昭. 68 例脑干海绵状血管瘤的临床特点及显微手术治疗效果分析[D]. 郑州大学, 2019.

[2] 孙季冬,刘翼,贺民,孙鸿,游潮. 脑干海绵状血管瘤的临川表现及预后分析[J]. West China Medical Journal, 2011, 355-358.

[3] 陈见清,包映辉,崔华,周正文,王勇. 显微外科治疗脑干海绵状血管瘤的研究进展[J]. 中国脑血管病杂志, 2015, 155-159.

[4] 宋国智,刘吉祥,常成陈建军,李海红,张钧,晁艳艳. Kawase’s 入路切除脑干海绵状血管瘤的临床研究[J]. Neural Injury and Functional Reconstruction, 2016, 512-523.

[5] 张力,王汉东,潘云曦,丁可,祝剑虹,茅磊. 脑干海绵状血管瘤的临床特点及显微手术治疗[J]. 中国脑血管病杂志, 2018, 543-548.

[6] 屠恩美, 杨杰. 半监督学习理论及其研究进展概述[J]. 上海交通大学学报, 2018, 1280-1291.

[7] Vapnik V. Statistical Learning Theory[M]. Wiley, New York, 1998.

[8] Blum A, Mitchell T. Combining Labeled and Unlabeled Data with Co-training[C]. Proceedings of the Eleventh Annual Conference on Computational Learning Theory. 1998, 92-100.

[9] Nigam K, Mccallum A K, Thrun S, et al. Text Classification from Labeled and Unlabeled Documents using EM[J]. Machine Learning, 2000, 39(2): 103-134.

[10] Rosenberg C, Hebert M, Schneiderman H. Semi-supervised Self-training of Object Detection Models[C]. Proceedings of the Seventh IEEE Work-shops on Application of Computer Vision, 2005.

[11] 帅平,李晓松,周晓华,刘玉萍.缺失数据统计处理方法的研究进展[J].中国卫生统计,2013,30(01):135-139+142.

[12] Breiman L . Random forest[J]. Machine Learning, 2001, 45:5-32.

[13] Stekhoven D J, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data[J]. Bioinformatics, 2012, 28(1):112-8.

[14] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. SMOTE:Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 321-357.

[15] 代冠华. 基于不平衡采样的分类预测模型研究[D]. 上海财经大学, 2020.

附录

附录 1：缺失数据分布可视化

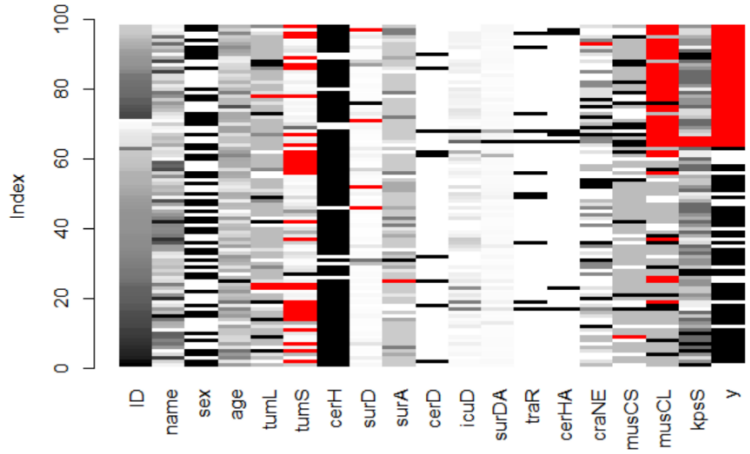


图 1 缺失数据分布可视化

附录 2：部分连续型变量箱线图

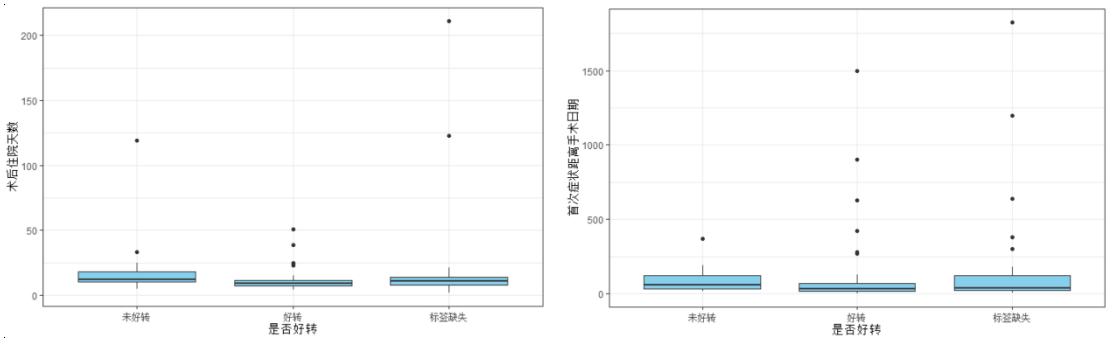


图 2 术后住院天数与首次症状距离手术日期分组箱线图

附录 3：基于 Kmeans-KNN 与随机森林两种填补方法的模型性能评价

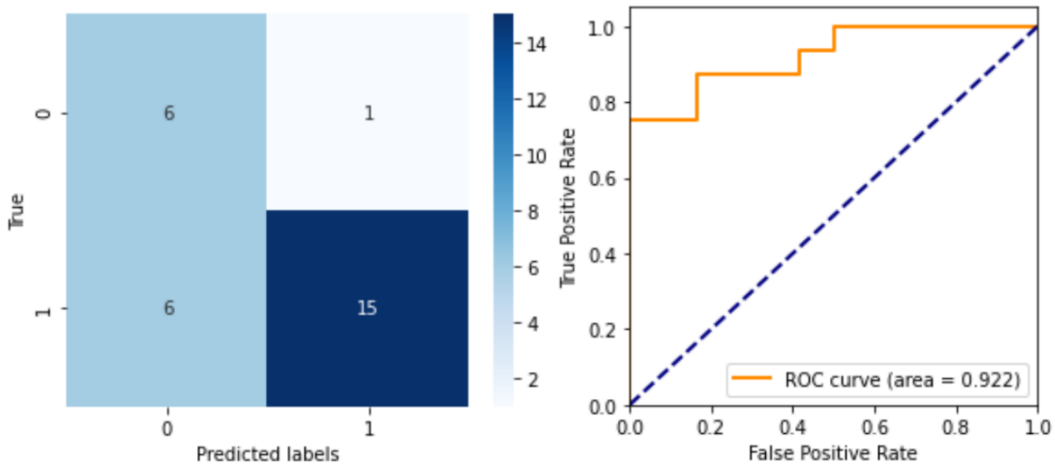


图 3 基于随机森林填补缺失值的建模效果图

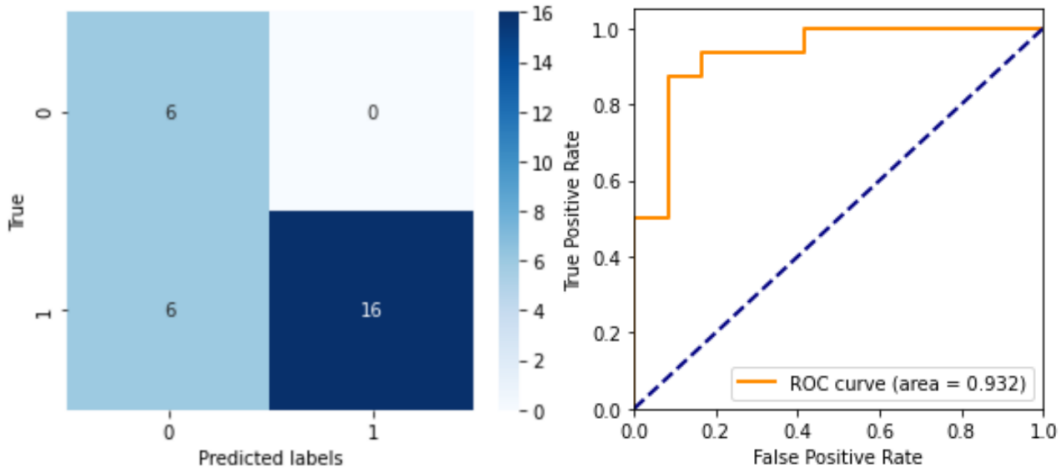


图 4 基于 Kmeans-KNN 填补缺失值的建模效果图

附录 4：相关缩写解释

相关缩写解释表

缩写	解释	缩写	解释
BSCM	脑干海绵状血管瘤	icuD	ICU 住院天数
sex	患者性别	surDA	手术后住院天数
age	患者年龄	traA	是否气切
tumL	肿瘤部位	cerHA	是否再出血
tumS	肿瘤大小	craNE	颅神经事件
cerH	是否出血	musCS	术后肌力变化
surD	首次症状距离手术日期	musCL	随访肌力变化
surA	手术入路	kpsS	术后 72hKPS
cerD	是否脑积水	y	因变量(随访 KPS-术前 KPS)

致谢

本篇论文完成在即，首先要感谢论文指导老师徐宇锋老师。徐老师对我们的论文研究方向做出了指导性的意见，在论文的撰写过程中，给予了许多的帮助，投入了许多的精力和心血。在此十分感谢徐老师的悉心指导。

其次，还要感谢中南大学湘雅医院心血管科室的侯伟学长，为我们提供了宝贵的原始数据，数据搜集不易，在此十分感谢侯学长的支持。

同时，还要感谢团队中的每一位成员，大家愉快合作，认真研究，不断沟通交流，最终才有了这篇论文的产生。

最后，我们要感谢培养我们长大的父母，在钻研课题遇到困难的时候，给予我们适当的理解，并在精神和物质上给予我们不断的支持。在此，我们向父母、家人以及朋友表示最真诚的感谢。