基于Python 3.12.4，pandas 2.2.2与scikit-learn 1.5.1库。

**数据清洗与缺失值处理**

原始数据集（data/raw/20250207 null\_sic\_dic.csv）包含10441条阴性样本、426条阳性样本，其数据结构特征如下：

* **非特征字段**：首列为患者ID（id）、倒数第二列为现行SIC诊断标准结果（sic），均为 1、结局指标（death），均不作为模型输入特征
* **标签定义**：最后一列dic为分类标签（0:阴性，1:阳性）

执行分阶段缺失值处理策略：

1. **特征筛选**：剔除缺失率超过50%的特征变量（本研究中无满足此条件的特征）
2. **样本筛选**：删除存在5个及以上缺失值的样本（共剔除3347条，保留8960条有效样本）
3. **缺失值插补**：
   * *连续性变量*：按标签组别（dic分组）计算列均值进行插补
   * *分类变量*：将缺失值统一编码为”0”，即表示不存在该种情况。  
     清洗后数据保存为data/清洗后.csv，其标签分布呈现显著不均衡性（阴性:阳性=8,531:428）。

**数据划分与标准化**

**分层数据分割**：  
采用分层抽样策略，按8:2比例将数据集划分为训练集（7,168条）与测试集（1,792条），确保两集合的标签分布与原始数据一致

**特征标准化**：

1. **Z-score标准化**：对连续型特征进行标准化处理，统一量纲，使得模型能均衡地考虑每个特征。
2. **步骤**：
   * 基于训练集计算均值与标准差参数
   * 将参数同步应用于测试集标准化，避免数据泄露  
     标准化后数据分别保存为data/train\_data\_standardized.csv与data/test\_data\_standardized.csv。