“数据仓库与数据挖掘”大作业

1. 任务背景

脓毒症是由感染引起的严重全身反应综合征，可能损害患者的多个器官，具有非常高的死亡率。如果能在早期根据脓毒症患者的生命体征信息，准确地识别出具有较高院内死亡风险的脓毒症患者，将有助于医护人员对患者的病情发展进行提前干预，降低脓毒症患者的院内死亡率。

生命体征天然地存在时序性，对脓毒症患者在过去一段时间内的生命体征时间序列进行建模分析，可以挖掘患者的历史病情发展规律，实现更准确地预测患者在未来某个时刻的死亡风险。

2. 任务描述

对脓毒症患者在过去 24 小时内的心率、呼吸率、平均动脉压以及血氧饱和度四项生命体征时序数据进行建模分析，预测患者在未来 6 小时后的死亡风险（死亡/存活）。

3. 数据描述

数据集记录患者在 24 小时内的生命体征数据，标签为观察窗口结束时的未来 6 个小时后患者是否死亡（图 1）。

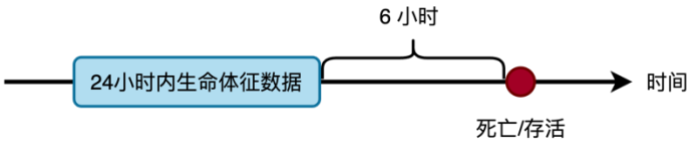

图 1 数据集数据记录示意图

训练集：project/dataset/train.csv

测试集：project/dataset/test.csv

字段名称	数据类型	字段描述	正常值范围
id	int	病人唯一标识	
time	str	时间戳	格式：YYYY-MM-DD HH:MM:SS
heartrate	str	心率	健康值参考范围：60-100
resprate	float	呼吸率	健康值参考范围：12-24
map	float	平均动脉压	健康值参考范围：70-105
o2sat	float	氧饱和度	健康值参考范围：95-100
label	int	是否院内死亡	0：存活 1：死亡

4. 作业要求

(1) **数据处理**：数据集可能存在数据缺失和数据异常的情况，因此需要对数据进行相关的预处理，包括但不限于数据清洗、数据修复、数据规范化等，可根据背景知识和数据形式自行决定。

(2) **模型算法**：使用至少 3 种时间序列建模分析方法完成任务，其中至少 1 种方法为自己实现，并比较自己的实现效果和机器学习库/参考文献中实现的差异。（鼓励对现有方法进行改进或创新）

(3) **评价指标**：根据任务特点选择合适的分类评价指标对模型进行评估，例如 Accuracy、Precision、Recall、F1、AUC、AUPRC 等，并说明选择原因。对预测结果绘制混淆矩阵和 ROC 曲线。

(4) **结果分析**：使用合适的方法对模型超参数进行选择，对比分析不同时间序列建模分析方法在该任务上的效果差异并分析原因。

5. 提交要求

(1) 本次作业分组完成，每组不超过 2 人。提交作业时由一人提交即可。

(2) 提交内容：

- 代码：所有实现作业要求所需的代码，语言不限。
- 实验报告：报告需要至少覆盖作业要求，中英文不限，不超过 6 页，以 pdf 格式提交。
- 作业分工说明：说明每位小组成员的学号、姓名以及在本次大作业中的工作内容。
- 以上内容打包成一个压缩文件上传，文件名格式：姓名 1_姓名 2_大作业.zip，如（张小明_李小华_大作业.zip）。

(3) 请把作业须在 2023 年 6 月 2 日 23:59:59（含）之前通过网络学堂提交。本次作业占课程总成绩 40%。