《云数据管理II -- 智能数据分析与决策支持》课程作业

1. 问题描述：建立分类模型预测糖尿病患者的治疗效果，并分析治疗效果受各因素的影响情况，治疗效果的评价指标为患者30天内是否再次入院。

根据patient\_nbr和readmitted来建立分类模型

1. 数据集：美国各地医院收集的部分临床记录，本实验仅保留糖尿病患者的相关记录，包括病人基本信息、既往病史、住院程序、入院诊断、生命体征、化验信息、用药记录、出院去向等信息。注：本实验所用数据集不包含任何可识别个人信息。
2. 任务说明：1）预处理阶段，包括缺失值处理，

同一个病人仅保留第一次入院记录（排除前次治疗对本次治疗效果的影响），

移除导致临终关怀或病人死亡的记录（通过discharge\_disposition\_id筛选）。

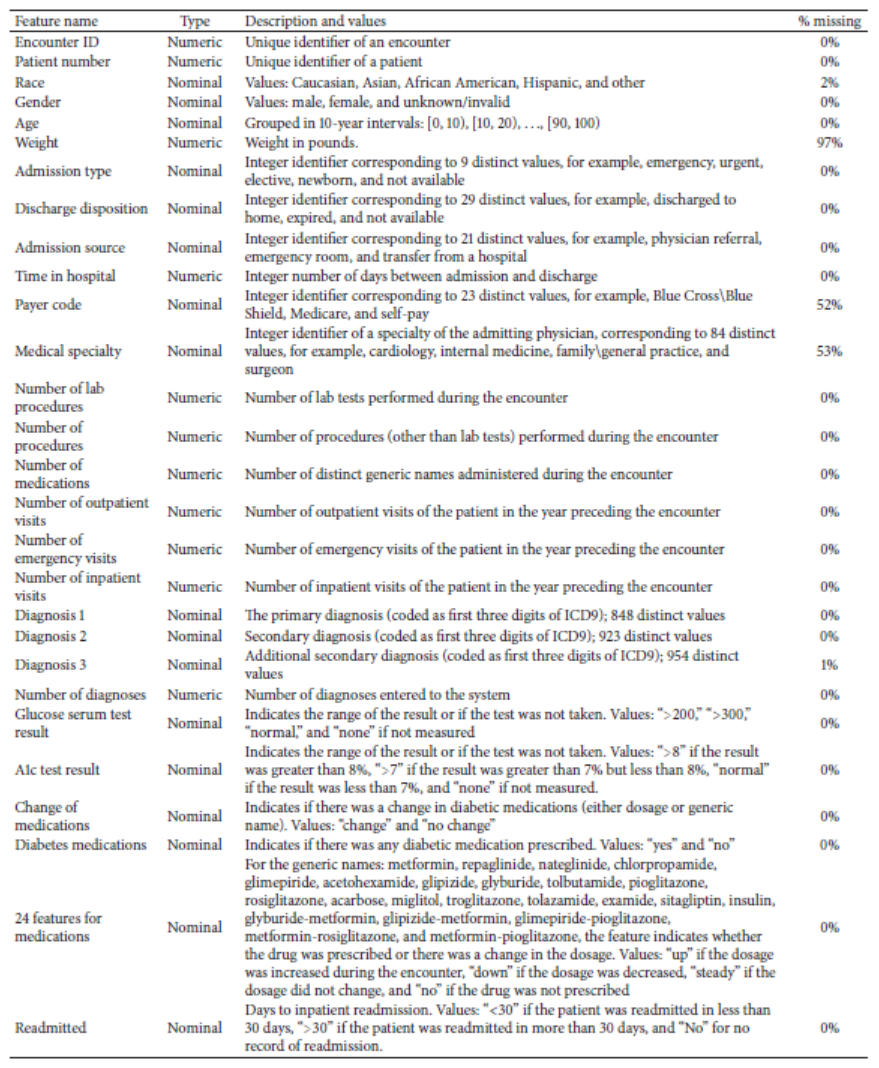
以图表形式统计余下病人年龄、种族、性别、入院类型、入院初诊（diag\_1）、出院去向等信息分布情况。

2）合理划分训练集和测试集，实现课程介绍的（但不限于）两种分类算法，用于预测患者的治疗效果，并比较分析各算法精度。

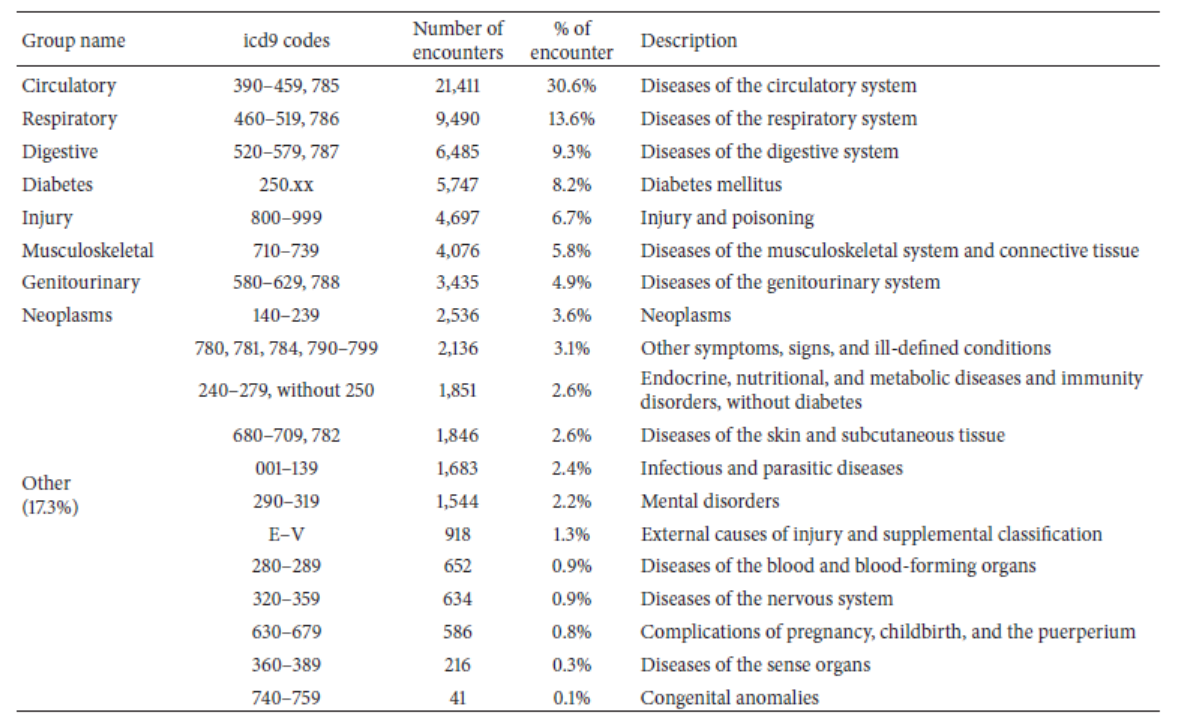
3）建立相关模型（logistic回归模型或其他）分析30天后再入院概率受各因素的影响情况。注：在建立模型之前可先做自变量筛选。

1. 作业说明：本次作业需独立完成，并在课程结束前进行答辩展示。
2. 提交内容：本实验需要在课程结束前进行答辩展示，需要提交源码及运行说明、实验报告（word或pdf，控制在10页以内）、答辩PPT（10分钟）。
3. 截止时间：请在2018年12月16日24:00前通过网络学堂提交。

附表一：数据集各字段说明信息



附表二：临床记录中诊断信息及其编号



医学小知识：1）血清葡萄糖检测：即血糖检测，在正常情况下，在胰岛素和其他激素的参与下，葡萄糖的分解代谢和生物合成处于动态平衡，血糖保持相对稳定。通过空腹血糖试验，可以确定是否存在高血糖、低血糖的情况。

2）A1C检测：糖化血红蛋白(HbA1c)是红细胞中的血红蛋白与血清中的糖类相结合的产物，它通过缓慢、持续及不可逆的糖化反应形成，其含量的多少取决于血糖浓度以及血糖与血红蛋白接触时间，而与抽血时间、患者是否空腹、是否使用胰岛素等因素无关。因此，HbA1c可有效地反映糖尿病患者过去1～2个月内血糖控制的情况。特别地，A1C检测常用于指导调整糖尿病治疗方案。