KDD Lab 作业 1

3/12 14:00~ 3/19 22:00 通过 blackboard 上传作业文件，本次作业不支持补交或缓交。

逾期不收！逾期不收！逾期不收！

第一次作业主要是让同学们学会熟练的使用 Python（Anaconda Jupyter Notebook），尤其是 pandas 库来协助读取，分析并清洗数据。

使用的数据为 HW1data.csv，该数据为医疗数据，包含以下内容：

PatientId：患者的唯一 ID 识别；

AppointmentID：预约 ID，也是唯一的；

Gender/Age/Neighbourhood：患者的一些基本信息；

ScheduledDay：预约医疗就诊的时间，精确到时分秒；

AppointmentDay：就诊日；

Scholarship/Hipertension/Diabetes/Alcoholism/Handcap：一些患者的相关信息

SMS_received：患者是否收到预约确认的短信

No-show：就诊日患者是否缺席

作业要求：按照下文中“数据预处理流程”清洗 HW1data.csv，并按照要求命名你的 ipynb 文件上传作业

命名要求：请按照“HW1+你的学号”的方式命名你的文件，假如学

号为 11956789，那么作业文件名为 HW1+11956789.ipynb

备注：

（1）为方便作业批改，尽量提交 ipynb 格式文件，提交前先清除所有输出，减少 ipynb 文件的大小

（2）为方便作业批改，请读取在同一个目录下的 HW1data.csv（比如在 read csv 时使用相对路径）

（3）请勿写入源文件 HW1data.csv！！

数据预处理流程（带*号的步骤是关键得分点&难点）：

1）成功读入数据，发现所有数据列都包含 10%左右的缺失

2）*通过下面一些数据补全的方法，将能够补全的内容进行补全：

a) 从原始数据中提取出所有的 PatientId 与其对应的 Gender/Age/Neighbourhood 信息，因为对于一个患者来说，只要知道 Id，性别年龄社区都是唯一并且可推断的

b) 回到最初的数据，仅去除掉 PatientID、ScheduledDay、AppointmentDay、SMS_received 和 No-show 的 NaN 值

c) 使用之前提取出来的 PatientId 与对应信息补全缺失的 Gender/Age/Neighbourhood 值

3）如果第二步去并没有对数据进行补全，则需要除掉 PatientID，Gender，Age，Neighbourhood，ScheduledDay，AppointmentDay，SMS_received 和 No-show 的 NaN 值

4）使用默认值 0 补全 Scholarship、Hipertension、Diabetes、Alcoholism

和 Handcap 信息，这里我们假设是这 5 个特征都可以使用 0 来补全，现实生活中对数据进行清洗操作的时候请注意根据对应的专业知识与理解来补全

5）从数据中去掉 PatientId 与 AppointmentID，因为这两个信息对模型训练没有用（本作业暂不涉及模型训练部分）

6）找出有问题的年龄（比如：Age = -1 的）并去除

7）计算出 ScheduledDay 和 AppointmentDay 的差距天数（Delta_Day），将此结果作为新的一列加入到数据中

8）*计算出 ScheduledDay 和 AppointmentDay 都分别是星期几（SDay_DOW，ADay_DOW），也将此结果加入到数据中

9）从数据中去掉 ScheduledDay 与 AppointmentDay，因为我们已经提取了相关信息了（Delta_Day，SDay_DOW，ADay_DOW）

作业计分（满分 10 分）

内容扣分点：

无法完成步骤 2：-1 分

无法完成步骤 8：-1 分

无法得到最终结果数据：-1 分

其他扣分点：

迟交：-10 分

不按照命名规则命名：-1 分

完全不注意备注信息：-1 分

程序无法执行/程序出现报错：-1 分