KDD Lab 作业 1

3/10 14:00 ~ 3/24 22:00 通过 blackboard 上传作业文件，作业不接受任何理由的补交或缓交，请同学们注意尽早完成并提交。截止时间为第六周周四晚上十点。Blackboard 上有 3 次提交机会，希望同学们提交前能好好检查完再上传，如果真的有需求重新上传，请联系助教赵奕丞处理。

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。数据中包括许多缺失，还有一些刻意加入的错误信息，需要同学们在做数据预处理的时候准确的识别出缺失与错误的内容并将数据清洗为整洁可用的数据。

使用的数据为 HW1data.csv，该数据为医疗数据，记录了许多患者（patient）在某所医院就医（appointment）的预约记录（schedule），以下是该数据的内容介绍：

PatientId：患者的唯一 ID 识别，针对某个患者，这个 ID 是唯一的；

AppointmentID：该医院的某一个预约 ID，也是唯一的；

Gender/Age/Neighbourhood：该患者的一些基本信息，性别，年龄，与该患者住处所属的街道（neighbourhood）；

ScheduledDay：该预约记录的时间（网上提前预约就医），精确到时分秒；

AppointmentDay：就诊日（实际该患者该出现在医院的时间）；

Scholarship/Hipertension/Diabetes/Alcoholism/Handcap：一些患者的就医相关个人信息；

SMS_received：患者是否收到预约成功的短信；

No-show：就诊日患者是否缺席（鸽了没）；

作业要求：按照下文中“数据预处理流程”中的 9 个步骤清洗 HW1data.csv，要求每一步都有明确的 ipynb print 记录，最好每个步骤都是 notebook 中一个独立的 cell，并明确的在程序代码中备注数据清洗的原因或者过程，最后按照要求命名你的 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，必须提交规定格式的 ipynb 文件，且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*号的步骤是难点）：

1）成功读入数据，发现所有数据列都包含缺失

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，性别年龄社区都应是唯一并且可推断的，且为了简化，我们可以假设人的年龄和社区都不会随时间变化（举例来说：PatientId 为 666 的张三，性别男，年龄 20，住在桃源街道；那么如果我们有一条 PatientId 为 666，性别为男，但缺少年龄和社区的记录，我们应该认为这是 20 岁，住在桃源街道的张三）

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（如果此处你有更好的补全方式，请备注说明原因）

5）从数据中去掉 PatientId 与 AppointmentID，因为这两个信息对未来的模型训练没有用（本作业不涉及模型训练）

6）找出有问题的年龄（比如：Age = -1 的）并去除（如果你有其他要去除的数据，请备注说明原因）

7）计算出 ScheduledDay 和 AppointmentDay 的差距天数(Delta_Day)，将此结果（预约和就诊之间的时间差）作为新的一列加入到数据中（这个数据是我们希望探讨，人们是否 Delta_Day 越大，越容易忘记就诊）

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

作业计分标准（满分 10 分，扣完为止）

内容扣分点：

某个步骤（共 9 个步骤）的数据清洗没有说明原因：-1 分

某个步骤（共 9 个步骤）没有 print 结果：-1 分

其他扣分点：

迟交：-10 分

作业文件不按照命名规则提交：-1 分

程序无法执行/程序出现无法继续运行的报错：-3 分

并没有提交 ipynb 文件，而是提交了 py 文件：第一次-5 分，第二次-8 分，第三次-10 分