# 基于Diabetes 130-US Hospitals for Years 1999-2008 Data Set数据集的糖尿病患者再入院预测与分析

姓名： 贾亚伟

学号：2021Z8017782015

专业： 电子信息

## 背景

## 糖尿病（Diabetes Mellitus）是最常见的慢性疾病之一，主要以高血糖为特征，其影响范围波及全球 4 亿多人，同时糖尿病具有明显的家族遗传特性，接近一半的糖尿病患者有家族遗传病史。糖尿病的病理原因主要包含两种情况，第一种是当胰腺无法产生充足的胰岛素（一种调节血糖或血葡萄糖的荷尔蒙）时引发的 1 型糖尿病（T1D），第二种是当所产生的胰岛素无法被人体有效地利用时引发的 2 型糖尿病（T2D）。T1D 通常被称为原发性糖尿病，这种糖尿病类型的发病机制通常是当胰腺中进行胰岛素分泌的 β 细胞受到损伤时，人体在短时间内没有充足的胰岛素供使用，从而导致血糖含量无法被及时降低至安全区间内，该过程也称作是胰岛β细胞的郎格罕氏（Langerhans）胰岛自身免疫性破坏。另一种更为常见的糖尿病类型的医学名称是非胰岛素依赖型糖尿病，简称 T2D。该类糖尿病通常由胰岛素抵抗或者是胰岛素分泌缺陷等因素引起，造成高血糖的直接原因是身体内的胰岛素没能得到有效利用。而引发 2 型糖尿病的主要原因通常包括生活方式、身体活动、饮食习惯和遗传等因素。

## 在中国过去三十多年的社会发展历史中，随着糖尿病患者人数不断上升，人们开始意识到这一普遍影响家庭生活和个人幸福的慢性疾病所带来的影响，但糖尿病患者的总数量仍然较二十多年前已经至少翻了一倍。

## 2017年全世界范围内糖尿病患病人群的数量将达到近四亿两千五百万。而根据近年的增长率预测到 2045 年全球的糖尿病患者将达到六亿两千九百万人，这个数将超过全世界总人口数量的近十分之一，这一惊人的数字毫无疑问需要引起我们的高度重视。

## 据研究表明，2013年至2018年期间，我国糖尿病的危险因素没有得到明显改善，部分危险因素甚至变得更为严重。例如，红肉摄入量过多的比例从32.6%增加到42.3%，身体活动不足率从16.0%增加到22.0%。

## 而根据我国判断肥胖的体重指数（BMI）标准，5年间我国肥胖患病率从14.1%上升到16.5%；中心性肥胖从31.6%增加到35.4%。到2018年，一半左右成年人处于超重或肥胖状态。因此，考虑到我国庞大的糖尿病前期人群数量以及还在上升的肥胖态势等因素，如果不加大糖尿病防控力度，未来我国糖尿病患病率可能会进一步增加。

## 因此，糖尿病患者的病情预测就变得十分有意义，本实验将在UCI machine learning提供的关于糖尿病的数据集的基础上探究糖尿病患者再入院的潜在风险特征因素。

## 数据概况

## 本实验使用的数据集为UCI machine learning提供的Diabetes 130-US Hospitals for Years 1999-2008 Data Set该数据集为对1999至2008年间不同经过医院实验室测试的十万名糖尿病患者进行描述的数据集，它经过临床专家的筛选后仅保留了50个最可能与糖尿病病情相关的属性，数据是代表弗吉尼亚联邦大学临床和转化研究中心提交的，该中心是 NIH CTSA 拨款 UL1 TR00058 的接受者和 CERNER 数据的接受者。该数据集代表了美国 130 家医院和综合交付网络的 10 年（1999-2008）糖尿病相关的临床护理数据。

## 该数据集包括代表患者和医院结果的50个特征。从数据库中提取满足以下标准的遭遇的信息。

## (1) 是住院会诊（入院）。

## (2) 这是一次糖尿病遭遇，即任何一种糖尿病都被输入系统作为诊断。

## (3)停留时间不少于1天，最多14天。

## (4) 在遭遇期间进行了实验室测试。

## (5) 会诊期间给予药物治疗。

## 数据包含以下属性：患者编号、种族、性别、年龄、入院类型、住院时间、入院医生的医学专业、进行的实验室检测次数、HbA1c 检测结果、诊断、用药次数、糖尿病药物、门诊人数住院前一年的住院、急诊等等。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集特点： | 多元 | 实例数量： | 100000 | 领域： | 生活类 |
| 属性特征： | 整数 | 属性数量： | 50 | 上传日期： | 2014.5.3 |
| 相关任务： | 分类/聚类 | 缺值： | 是 | 浏览次数： | 377464 |

## 表1.数据集的相关信息

## 该数据集用在预测糖尿病患者是否会出院后再入院方面有好的效果，因此本实验也使用该数据集做相关的机器学习预测分类任务。

## 方法

## 本实验方法分为两个部分，第一部分为数据预处理阶段对于数据集中数据的清洗即空缺值的预测填充；第二部分为根据清洗后的数据对糖尿病患者再入院的预测

## 实现步骤

## 首先，收集到实验使用的数据集后的第一个任务就是将该数据集导入到项目中去，然后查看数据集的shape以及其数据集的描述。可以了解到共有101766个患者记录，每个记录具有50个功能。由于该数据及包含了同一患者的多次记录，而本实验的目的只为了研究糖尿病患者30天内再入院的可能，因此我们需要去除掉同一用户的重复信息，同时该数据集还包含了已经死亡以及临终关怀的患者，为更有利于研究我们也要把这一部分患者去除掉。

## Image for post数据清洗的另一大任务就是去除过量含有空值的属性和将部分空值进行合理的赋值。

## 图1.各属性空值的分布情况

## 需要根据空值的分布图来进行部分属性的取舍，由于体重的空值已经占了96%，属于高度数据缺失，因此将该列去处。并且支付人代码和医学专业这两个属性的空值量也比较大43%和48%，但可以通过使用分布预测与归因技术来填补其空缺值，因此将其保留。

## 在单变量分析环节，首先观察了种族的分布，发现种族分布主要是以白种人为主其次是非洲裔美国人和亚裔，因此将该属性中占比2.7%的缺失值替换为白种人；在性别分布这一属性中，其中有3个性别值是未知的，选择把这些行给删除掉，从性别分布中可以看出女性人数多于男性人数，但差异很小；在年龄分布中发现，年龄小于40岁的患者与年龄大于40岁的患者相比数量较少，患者主要分布在70—80年龄段内，为了便于后续研究将患者年龄分为0—30、30—60、60—100三类；在入院类型分布中，可以看出大部分患者都是因急诊而入院；在出院处置类型分布中，可以看出大部分患者都回了家，又因为原本的分类过于糅杂，因此在进行重新归类将原本的29类归为了8类；再入院来源分布中可以看出大多是来自于急诊，其次就是转诊，后又将原本的26类入院类型归纳为8类；在住院时间分布中可以看出，患者的住院时间分为1天到14天不等，患者平均停留4天，大部分患者停留3-4天，患者很少停留超过12天；在遭遇期间进行的实验的数量分布中可以看出，实验者的平均实验次数为43次；在遭遇期间执行的程序（实验室测试除外）的数量分布中可以看出，大部分患者没有再做相关的实验；在患者的摄入药品种类分布中可以看出，大多数患者平均获得 16 种药物，只有 7 名患者服用了 70 多种药物；在患者就诊前一年的就诊次数中可以看出，大多数患者没有任何门诊就诊记录，门诊次数超过 15 次的患者非常少；在患者入院前一年的急诊次数分布中可以看出，大部分患者都没有急诊记录；在患者入院前一年的入院记录分布中可以看出，大部分患者就诊前都没有住院就诊记录；在三次诊断记录中，将诊断结果分为9类。