天池精准医疗大赛－糖尿病风险预测

- 葛传斌　2120171404
- 殷赫　　2220170871
- 李玥儿　2120171411
- 秦娜　　2120171418

一、引言

进入 21 世纪，生命科学特别是基因科技已经广泛而且深刻影响到每个人的健康生活，于此同时，科学家们借助基因科技史无前例的用一种全新的视角解读生命和探究疾病本质。人工智能（AI）能够处理分析海量医疗健康数据，通过认知分析获取洞察，服务于政府、健康医疗机构、制药企业及患者，实现个性化，可以循证的智慧医疗，推动创新，实现价值。

心血管病、糖尿病等慢性疾病，每年导致的死亡人数占总死亡人数的 80%，每年用于慢病医疗费用占中国公共医疗卫生支出的比例超过 13%。糖尿病是一组以高血糖为特征的代谢性疾病，糖尿病时长期存在的高血糖，导致各种组织，特别是眼、肾、心脏、血管、神经的慢性损害、功能障碍，现已成为目前较为普遍的疾病。我国是世界上糖尿病患者最多的国家，病人达到 1.1 亿，每年有 130 万人死于糖尿病及其相关疾病。每年用于糖尿病的医疗费用占中国公共医疗卫生支出的比例超过 13%，超过 3000 亿元。

二、问题描述与评价指标

1、　问题描述

作为一种常见慢性疾病，糖尿病目前无法根治，但却能通过科学有效的干预、预防和治疗，来降低发病率和提高患者的生活质量。通过糖尿病人的临床数据和体检指标来预测人群的糖尿病程度，以血糖浓度为指标，设计高精度，高效，且解释性强的算法来进行糖尿病的预测，即用人工智能的方法和思想处理、分析、解读和应用糖尿病相关大数据，辅助糖尿病的预测。

2、　数据集描述

数据来自天池大数据竞赛，包括训练文件 d_train.csv 和测试文件 d_test.csv，每个文件第一行是字段名，之后每一行代表一个个体。文件共包含 42 个字段，包含数值型、字符型、日期型等众多数据类型，部分字段内容在部分人群中有缺失，其中第一列为个体 ID 号。训练文件的最后一列为标签列，即需

要预测的目标血糖值。

3、 评价指标

将预测的血糖值于个体实际检测到的血糖结果进行对比，以均方误差为评价指标，结果越小越好，均方误差计算公式如下：

$$f = \frac{1}{2m}\sum_{i=1}^{m}(y'(i) - y(i))^2$$

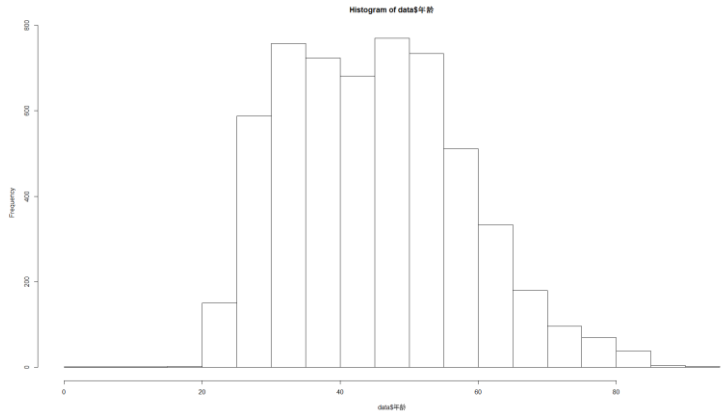
其中 m 为总人数，$y'(i)$为选手预测的第 i 个人的血糖值，$y(i)$为第 i 个人的实际血糖检测值。

三、模型建立

1、 数据探索性分析

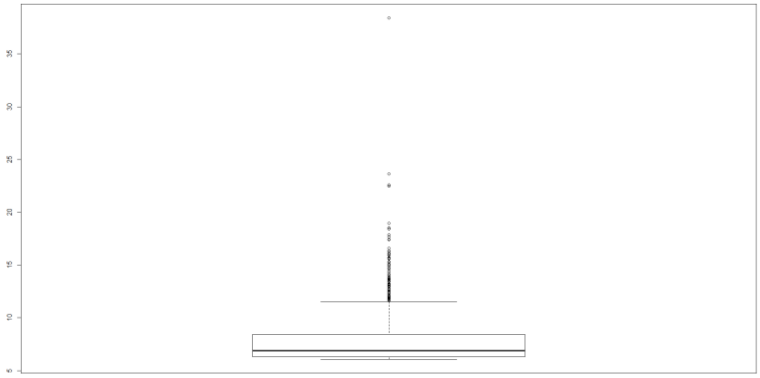
训练数据集包含有 42 个字段（包括血糖值在内），去除日期，ID，和若干个字符型字段，剩下了包括血糖在内的 39 个数值型字段，除了年龄之外均代表了各种临床生化检测指标，我们以这 39 个数值属性建立了我们建模需要的训练数据集。训练数据集包含有 5641 条病人的数据，首先观察数据的缺失情况，有 15 个字段的缺失比例为 20%左右，乙肝相关的 5 个字段的缺失比例较大，为 76%，剩余字段的缺失比例在 5%以内。

病人的年龄分布图如下所示，



从医学角度而言，对于其它的生化检测指标，均存在一个正常值范围，在这个范围之内的值都可以认为是正常的，过高过低的数值才被认为是异常的。我们查询了血糖的正常值范围，空腹血糖的正常范围是 3.61～6.05，对数据进行分析发现，几乎没有在 3.61 以下的血糖值，而高于 6.05 的血糖值有较多，而且跨度非常大，最大值甚至达到了 38 左右。而血糖值过高正是我们判断糖尿病的标准之一，因此这个问题需要对血糖值进行预测。我们对高于 6.05 的血糖值单独提

出绘制了箱形图如下所示，



最小值	第一四分位数	中位数	均值	第三四分位数	最大值
6.06	6.312	6.870	7.882	8.410	38.430

我们可以看到，大部分血糖值都只是略高于正常值范围，考虑到人体的自我调节功能，其实略微偏高的血糖值都是可以认为是正常的血糖值，我们认为血糖值大于第三四分位数的患者才是糖尿病的高危患者。

2、 数据预处理

数据预处理主要是针对数据中缺失值的处理以及数据的归一化。

对于除了乙肝 5 项缺失比例较多之外的 32 个字段，采用均值对缺失值进行填充；对于缺失比例高达 76%的 5 项乙肝相关字段，有两种方案，一是直接删掉相应的 5 个字段，二是不进行填充，利用一些能够存在缺失情况的数据进行建模的预测方法；考虑到乙肝相关的 5 个指标在医学上是认为与糖尿病有密切联系的，而缺失的比例过大，没有合适的方法进行缺失值的填充，我们采用了第二种方案，而我们后面采用的预测模型建立的算法正好也能够对存在缺失情况的数据集进行预测建模，我们也对删除和保留这几项字段的结果进行了对比。

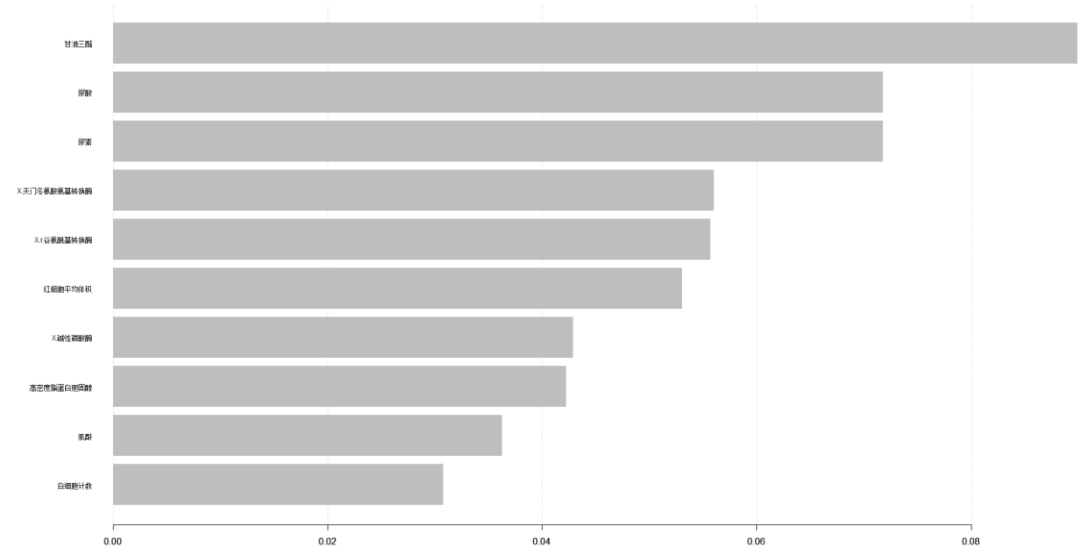
数据的归一化处理上，我们也采用了两种方法，一是基于数据最大最小值的 min-max 方法，二是基于数据服从正态分布假设的归一化方法。

3、 预测模型建立

在完成了数据的预处理之后，我们采用 XGBoost 方法进行了血糖值预测模型的建立。XGBoost 是一种集成学习方法，将多个分类回归树进行组合，以达到更强的泛化能力和更好效果。

针对两种对于 5 项乙肝指标的不同处理方式分别建模进行了对比，同时也以两种不同的归一化方式进行了建模，共四个模型，计算评价指标均方误差，四个模型的表现差异不大，模型得到的变量重要性如下图所示，只取了前 10 个字段

进行展示。



四、分析与总结

四个模型的均方误差分别为 1.01，1.02，0.96，0.89；最终在以数据服从正态分布的假设下的归一化处理与保留 5 项乙肝指标的情况下达到了最小的均方误差 0.89。在已知使用了类似方法的前提下，与比赛排名靠前的队伍相差还是较多，单纯对数据缺失值处理，归一化处理，以及设置模型的参数，对模型最终的预测结果精度的提升不大，需要考虑在特征工程上进行探索，尤其是对于医学上的数据，往往是有一个正常值的范围，而这个范围却没有明确的界限，这是在对医学数据处理上的一个疑难点，而且加上个体化的差异，数据需要更加科学合理的处理方法，才能对医学数据挖掘取得一个较好的结果。