2023 年中国地质大学（武汉）数学建模培训模拟竞赛题目

（请先阅读“全国大学生数学建模竞赛论文格式规范”）

B 题关于糖尿病风险预测

糖尿病是一种代谢性疾病，它的特征是患者的血糖长期高于标准值，当胰腺产生不了足够的胰岛素或者人体无法有效地利用所产生的胰岛素时，就会出现糖尿病。血糖正常值是指人空腹的时候血糖值在 3.9～6.1 毫摩尔/升，但是作为判断是否有高血糖，一 般是对人体进行两次重复测量，血糖值大于 6.7 毫摩尔/升即可诊断为糖尿病。临床表现为频尿、容易口渴、容易饥饿。同时伴随并发症：心血管疾病、非酮症之超渗透压的昏迷、糖尿病酮症酸中毒、中风、慢性肾脏病、足部溃疡等。目前糖尿病种类主要分为：1 型糖尿病、2 型糖尿病、妊娠糖尿病和其他类型糖尿病。糖尿病作为一种常见慢性疾病，目前无法根治，需要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。2021 年 IDF 发布的《全球糖尿病地图(第 10 版)》数据显示，全球成年糖尿病患者人数达到 5.37 亿（10.5%），约十分之一的成年人受到影响。相比 2019 年，糖尿病患者 增加了 7400 万，增幅达 16%，突显出全球糖尿病患病率的惊人增长。过去的 10 年间（2011 年~2021 年），我国糖尿病患者人数由 9000 万增加至 1 亿 4000 万，增幅达 56%，其中约 7283 万名患者尚未被确诊，比例高达 51.7%。

附件 1~2 分别给出了有血糖值的检测数据和无血糖值的检测数据，文件数据的部分特征名已做脱敏处理，包含年龄、性别、各项体检数据等 42 个监测指标(详见表 1)，包含数值型、字符型、日期型等数据类型。请你们团队根据实际和附件中的数据信息，通过建立数学模型研究主要解决下列问题：

1. 结合附件**1** 的检测数据信息，对样本数据进行预处理，并作相关的统计分析。

2. 在问题 **1** 的基础上，通过降维的方法从 42 个检测指标中筛选出主要变量指标，使之尽可能具有代表性、独立性(为了应用方便，建议降维后的主要指标在10个以下)，并请详细说明建模主要变量的筛选过程及其合理性。

3. 在问题2基础上，采用上述样本和建模的主要变量，通过数据挖掘技术建立血糖值预测模型，并进行模型验证。

4. 在问题3基础上，利用该模型对附件2的检测数据的血糖值进行预测，并对糖尿病风险进行说明。

附件1：有血糖值的检测数据.csv
附件2：无血糖值的检测数据.csv

表1　特征名称说明

特征名称	特征名称
id	乙肝e 抗原
性别	乙肝e 抗体
年龄	乙肝核心抗体
体检日期	白细胞计数
*天门冬氨酸氨基转换酶	红细胞计数
*丙氨酸氨基转换酶	血红蛋白
*碱性磷酸酶	红细胞压积
*r-谷氨酰基转换酶	红细胞平均体积
*总蛋白	红细胞平均血红蛋白量
白蛋白	红细胞平均血红蛋白浓度
*球蛋白	红细胞体积分布宽度
白球比例	血小板计数
甘油三酯	血小板平均体积
总胆固醇	血小板体积分布宽度
高密度脂蛋白胆固醇	血小板比积
低密度脂蛋白胆固醇	中性粒细胞%
尿素	淋巴细胞%
肌酐	单核细胞%
尿酸	嗜酸细胞%
乙肝表面抗原	嗜碱细胞%
乙肝表面抗体	血糖