基于机器学习的糖尿病预测

课题想法来源于与我刚结束的[阿里天池大赛](https://tianchi.aliyun.com/competition/information.htm?spm=5176.100067.5678.2.4a6372b3J21C63&raceId=231638)，在2522支队伍中取得了第31名的成绩。

**一、关于课题**

中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。传统的血糖监测方法需要抽血采样，而随着机器学习和大数据医疗的兴起，基于患者医疗体检的数据预测血糖有了一定的可行性。

本文的研究目的在于利用机器学习预测患者的血糖值，为医生诊断提供参考，减少患者不必要的抽血检查。本文主要的工作包括以下几个方面：

（1）提出了多角度的体检特征。对患者数据进行预处理、均衡化，进行特征工程，更全面地描述患者身体指标。

（2）研究了多种回归预测模型，探究最优的参数配置。对包括传统模型如支持向量机、随机森林、XGBoost，最新模型如CatBoot 在内的多种模型及其参数进行了研究，通过多种指标比较了模型的性能。

（3）模型融合进行血糖预测，通过stacking，boosting等方法进行了模型融合，进一步提高预测的准确度。

（4）特征重要性排序，在训练的过程中记录特征重要性，确定较重要的几个特征，为医生诊断提供额外的参考。

**二、关于该课题的思考**

1、首先该课题具有较好的现实意义和研究价值，机器学习在医疗方面的应用是一个新兴领域，相关研究较少，相对容易出成果。

2、其次该课题对学生具有一定的挑战性，运用到了机器学习的完整流程，如特征工程、模型训练、参数调整、模型融合等。经过该课题的训练可以让学生对机器学习具有一定的了解。

3、该课题我也是刚刚做完，还比较熟悉，可以给他们一些指导，避免弯路。此外遇到的问题可以一起探讨，相信在这个过程中学生们也能学到一些东西，如搜集资料、debug、代码理解等。

**三、关于如何指导**

1、了解学生个人情况，如编程语言、代码功底、机器学习了解程度等，然后制定辅导计划；

2、安排任务收集相关论文，Github实现等，并对其进行把关，做好辅导和建议；

3、指导实验，把控实验进度；

4、对论文写作过程中的问题给出指导。

简而言之，一切以学生为主，让学生有所收获。具体的计划要和工作人员进一步沟通，明确此次指导的具体目标、任务之后才能出炉。我也准备了部分文献和代码，可以为学生们提供参考。

最后，相信我有能力也有热情去完成此次指导，并能和学生们成为好朋友。