如何打一个数据挖掘比赛（入门版）

贡献者名单

代码贡献

牧小熊、司玉鑫

代码测评

潘姝宇、王振东、骆秀韬

贡献组织名单

厦门大学WISERCLUB-竞赛部

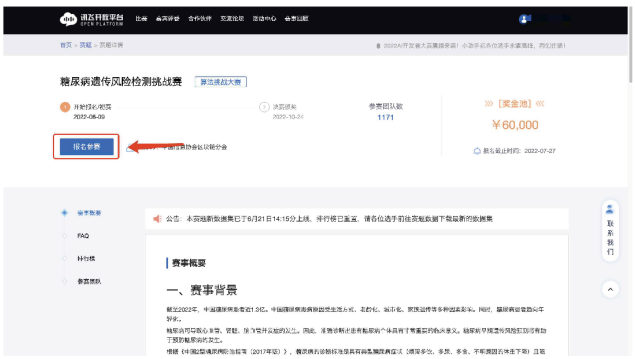
这是一份简易的竞赛教程，我们的目的是帮助同学们迈出AI训练大师之路的第一步。数据挖掘中会有很多需要学习的地方，建议入门的同学可以暂时不用着急去弄各个代码的原理，先跑通代码，然后看代码中涉及的知识点去查询相关资料进行学习，这样能让你学习更加有目标性，也容易找到学习的乐趣。千里之行，始于足下，从这里，开启你的AI学习之旅吧！

1.准备步骤

平台注册与比赛报名

1. 点击赛事链接：https://challenge.xfyun.cn/topic/info?type=diabetes&ch=ds22-dw-wd06
2. 注册（记得填写个人信息）

3. 点击报名参赛，显示成功报名



数据下载

1. 数据获取
    - 官网下载数据：下载数据及实名验证
    - 请把数据文件和代码文件放在同一个文件夹下，保证正常运行

环境配置

- python环境的搭建请参考
  - Mac设备：Mac上安装Anaconda最全教程
  - Windows设备：Anaconda超详细安装教程(Windows环境下)
  - Anaconda的介绍与安装，jupyter notebook安装和使用

2.实践思路

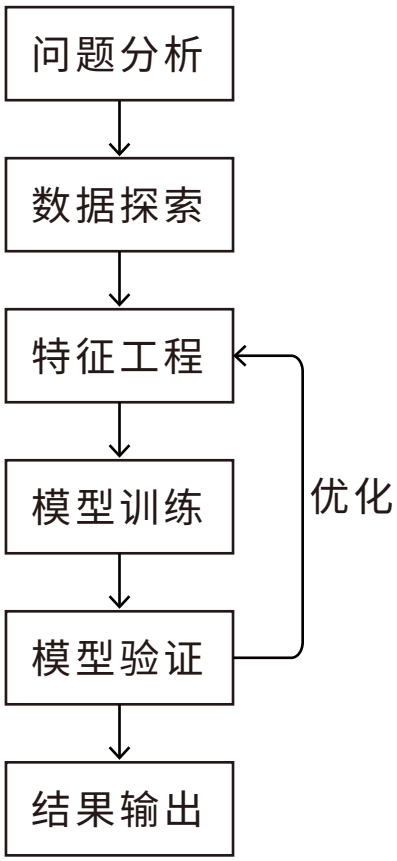
思路讲解

本次比赛是一个数据挖掘赛，需要选手通过训练集数据构建模型，然后对验证集数据进行预测，预测结果进行提交。

本题的任务是构建一种模型，该模型能够根据患者的测试数据来预测这个患者是否患有糖尿病。这种类型的任务是典型的二分类问题（患有糖尿病/不患有糖尿病），模型的预测输出为0或1（患有糖尿病：1，未患有糖尿病：0）。

机器学习中，关于分类任务我们一般会想到逻辑回归、决策树等算法，在这个Baseline中，我们尝试使用决策树来构建我们的模型。

我们在解决机器学习问题时，一般会遵循以下流程：



代码实现

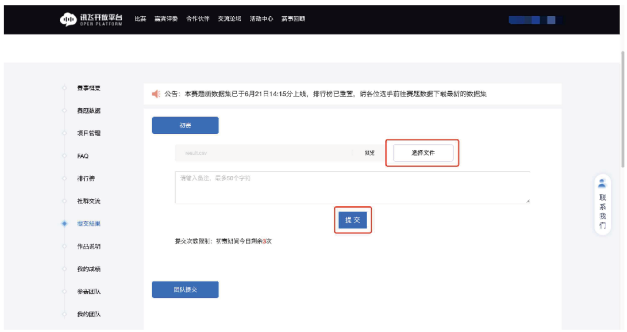
```
#安装相关依赖库 如果是windows系统，cmd命令框中输入pip安装，参考上述环境配置
#!pip install sklearn
#!pip install pandas
#-------------------------------------------------
#导入库
#---------------数据探索---------------
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
#数据预处理
data1=pd.read_csv('比赛训练集.csv',encoding='gbk')
data2=pd.read_csv('比赛测试集.csv',encoding='gbk')
#label标记为-1
data2['患有糖尿病标识']=-1
#训练集和测试机合并
data=pd.concat([data1,data2],axis=0,ignore_index=True)
#将舒张压特征中的缺失值填充为-1
data['舒张压']=data['舒张压'].fillna(-1)

#---------------特征工程---------------
"""
将出生年份换算成年龄
"""
data['出生年份']=2022-data['出生年份'] #换成年龄


"""
人体的成人体重指数正常值是在18.5-24之间
低于18.5是体重指数过轻
在24-27之间是体重超重
27以上考虑是肥胖
高于32了就是非常的肥胖。
"""
def BMI(a):
    if a<18.5:
        return 0
    elif 18.5<=a<=24:
        return 1
    elif 24<a<=27:
        return 2
    elif 27<a<=32:
        return 3
    else:
        return 4

data['BMI']=data['体重指数'].apply(BMI)
```

```
46
47  #糖尿病家族史
48  """
49  无记录
50  叔叔或者姑姑有一方患有糖尿病/叔叔或姑姑有一方患有糖尿病
51  父母有一方患有糖尿病
52  """
53  def FHOD(a):
54      if a=='无记录':
55          return 0
56      elif a=='叔叔或者姑姑有一方患有糖尿病' or a=='叔叔或姑姑有一方患有糖尿病':
57          return 1
58      else:
59          return 2
60
61
62  data['糖尿病家族史']=data['糖尿病家族史'].apply(FHOD)
63  """
64  舒张压范围为60-90
65  """
66  def DBP(a):
67      if a<60:
68          return 0
69      elif 60<=a<=90:
70          return 1
71      elif a>90:
72          return 2
73      else:
74          return a
75  data['DBP']=data['舒张压'].apply(DBP)
76
77
78  #---------------------------------
79  #将处理好的特征工程分为训练集和测试集，其中训练集是用来训练模型，测试集用来评估模型准确度
80  #其中编号和患者是否得糖尿病没有任何联系，属于无关特征予以删除
81  train=data[data['患有糖尿病标识'] !=-1]
82  test=data[data['患有糖尿病标识'] ==-1]
83  train_label=train['患有糖尿病标识']
84  train=train.drop(['编号','患有糖尿病标识'],axis=1)
85  test=test.drop(['编号','患有糖尿病标识'],axis=1)
86
87  #----------------模型训练----------------
88  model = DecisionTreeClassifier()
89  model.fit(train, train_label)
90  y_pre=model.predict(test)
91  y_pre
92
93  #----------------结果输出----------------
94  result=pd.read_csv('提交示例.csv')
```

```
result['label']=y_pre
result.to_csv('result-de.csv',index=False)
```

结果提交

- 在提交结果处提交，提交`预测结果.csv`，查看自己的成绩排名。



选择刚才生成的result.csv点击提交

点击我的成绩查看结果

3.学习提升

数据挖掘流程及参考学习资料：xunfei_demo.ipynb（用jupyter notebook打开）

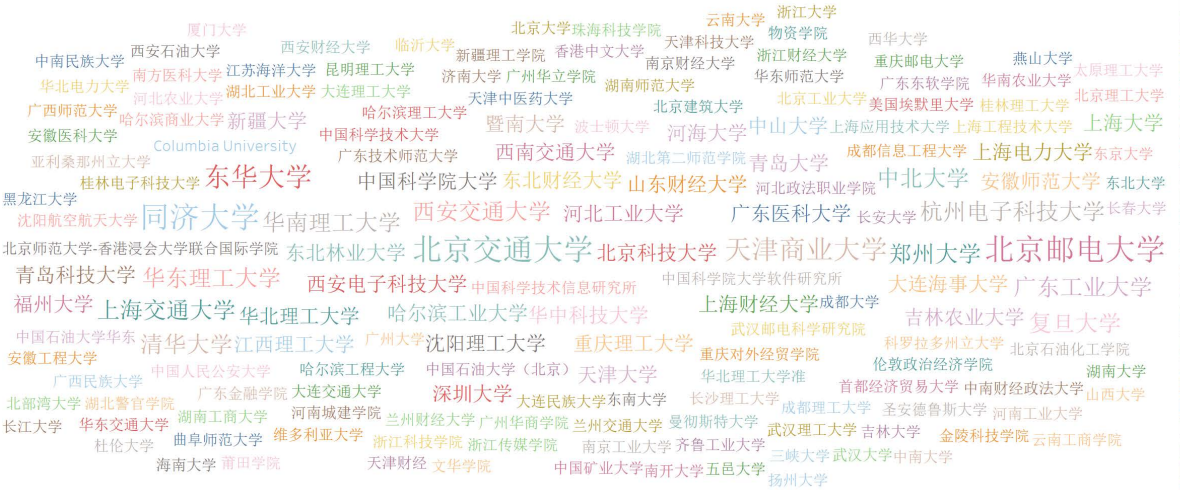
4.提问

如您遇到更多问题，可在Q&A文档中记录，我们会跟进反馈

https://docs.qq.com/sheet/DUEVNbENIT1ZiZ2t2?tab=BB08J2

评测贡献者

评测贡献高校


厦门大学
北京大学珠海科技学院
云南大学
浙江大学
物资学院
西华大学
中南民族大学
西安石油大学
西安财经大学
临沂大学
新疆理工学院
香港中文大学
天津科技大学
南京财经大学
浙江财经大学
重庆邮电大学
燕山大学
南方医科大学
江苏海洋大学
昆明理工大学
济南大学
广州华立学院
华东师范大学
广东东软学院
华南农业大学
太原理工大学
华北电力大学
河北农业大学
湖北工业大学
大连理工大学
天津中医药大学
湖南师范大学
北京工业大学
美国埃默里大学
桂林理工大学
北京理工大学
广西师范大学
哈尔滨商业大学
新疆大学
哈尔滨理工大学
北京建筑大学
暨南大学
波士顿大学
河海大学
中山大学
上海应用技术大学
上海工程技术大学
上海大学
安徽医科大学
Columbia University
中国科学技术大学
广东技术师范大学
西南交通大学
湖北第二师范学院
青岛大学
成都信息工程大学
上海电力大学
东京大学
亚利桑那州立大学
桂林电子科技大学
东华大学
中国科学院大学
东北财经大学
山东财经大学
河北政法职业学院
中北大学
安徽师范大学
东北大学
黑龙江大学
沈阳航空航天大学
同济大学
华南理工大学
西安交通大学
河北工业大学
广东医科大学
长安大学
杭州电子科技大学
长春大学
北京师范大学-香港浸会大学联合国际学院
东北林业大学
北京交通大学
北京科技大学
天津商业大学
郑州大学
北京邮电大学
青岛科技大学
华东理工大学
西安电子科技大学
中国科学技术信息研究所
中国科学院大学软件研究所
大连海事大学
广东工业大学
福州大学
上海交通大学
华北理工大学
哈尔滨工业大学
华中科技大学
上海财经大学
成都大学
吉林农业大学
复旦大学
中国石油大学华东
清华大学
江西理工大学
广州大学
沈阳理工大学
重庆理工大学
武汉邮电科学研究院
科罗拉多州立大学
北京石油化工学院
安徽工程大学
重庆对外经贸学院
伦敦政治经济学院
湖南大学
广西民族大学
中国人民公安大学
哈尔滨工程大学
中国石油大学（北京）
天津大学
华北理工大学淮
首都经济贸易大学
中南财经政法大学
山西大学
北部湾大学
湖北警官学院
广东金融学院
大连交通大学
深圳大学
大连民族大学
东南大学
长沙理工大学
成都理工大学
圣安德鲁斯大学
河南工业大学
长江大学
华东交通大学
湖南工商大学
河南城建学院
兰州财经大学
广州华商学院
兰州交通大学
曼彻斯特大学
武汉理工大学
吉林大学
金陵科技学院
云南工商学院
杜伦大学
曲阜师范大学
维多利亚大学
浙江科技学院
浙江传媒学院
南京工业大学
齐鲁工业大学
三峡大学
武汉大学
中南大学
海南大学
莆田学院
天津财经
文华学院
中国矿业大学
南开大学
五邑大学
扬州大学