****

****

**QG中期考核详细报告书**

**题    目**

**学   院**

**专 业**

**年级班别**

**学 号**

**学生姓名**

**年 月 日**

1、数据清洗：

首先先调用numpy，pandas，missingno库来处理数据。将csv中的文件导入并构造datafram类型的数据dataset，用describe()方法查看数据集的数据，再调用matrix()方法查看数据中的空值，可以看出，其他特征的缺失量并不大，所以选择直接删去有其他特征缺失的数据，而年龄缺失很多则选择填充平均值数据。先删除四项无关数据，再将年龄填入平均值，并用dropna（）方法去掉含有空值的行。

2、数据处理：

清洗后的数据要进行简单的数据处理，将非数值的特征用特定的方式转化成数值，便于之后模型的数学计算。将只有两种可能，且相互没有数值联系的性别，测试类型，是否残疾这三个特征用0，1进行编码。将数值间有大小关系的学历，难度和测试程序用0，1，2，3，4等进行编码。特征编码后用train\_test\_split()将数据分成训练集和测试集后再进行特征缩放和归一化处理。

3、模型：

显然，这是一个分类问题，逻辑回归虽然叫回归却是一个分类算法，逻辑回归实现简单，广泛用于二分法类的问题上，对于初学者的我无疑是个不错的选择。

选择好模型之后，调用sklearn库中关于逻辑回归的内容，初始化逻辑回归类，将处理后的分出的训练集传递过去，并得到预测结果，调用accuracy\_score（）得到测试的准确率，之后再交叉验证综合准确率。

4、实现：

将测试集读入，进行与训练集类似的数据处理，并调用predict（）函数得到结果，再将结果输出到本地中。