**广州大学学生实验报告**

**开课学院及实验室：**计算机科学与网络工程学院/电子信息楼  **2020 年 10月 1日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学 院** | 计算机科学与网络工程学院 | **年级/专业/班** |  | **姓名** |  | **学号** |  |
| **实验课程名称** | **机器学习与数据挖掘实验** | | | | | **成绩** |  |
| **实验项目名称** | **实验1** | | | | | **指导老师** | 彭伟龙 |

(\*\*\*（1）报告只能为文字和图片；（2）实验项目名称不能有任何改动；（3）必须填写正确的姓名和学号；否则会影响最后成绩的统计，后果自负\*\*\*)

1. **组员：**

**汤韬1806300183，李炜煜1806300182，张洪鑫1806300003**

1. **实验目的及要求**

1. 掌握对数据集成、清洗概念的理解

2. 掌握数据清洗，样本数据构建基础方法

3. 掌握基本的统计方法

4. 按照既定格式书写实验报告

**三、实验设备与平台**

1. 实验设备：计算机；

2. 平台：Windows 10操作系统

**四、实验内容**

**题目**：广州大学某班有同学100人，现要从两个数据源汇总学生数据。第一个数据源在数据库中，第二个数据源在txt文件中，两个数据源课程存在缺失、冗余和不一致性，请用C/C++/Java程序实现对两个数据源的一致性合并以及每个学生样本的数值量化。

数据库表：ID (int), 姓名(string), 家乡(string:限定为Beijing / Guangzhou / Shenzhen / Shanghai), 性别（string:boy/girl）、身高（float:单位是cm)）、课程1成绩（float）、课程2成绩（float）、...、课程10成绩(float)、体能测试成绩（string：bad/general/good/excellent）；其中课程1-课程5为百分制，课程6-课程10为十分制。

txt文件：ID(string：6位学号)，性别（string:male/female）、身高（string:单位是m)）、课程1成绩（string）、课程2成绩（string）、...、课程10成绩(string)、体能测试成绩（string：差/一般/良好/优秀）；其中课程1-课程5为百分制，课程6-课程10为十分制。

两个数据源合并后读入内存，并统计：

1. 学生中家乡在Beijing的所有课程的平均成绩。
2. 学生中家乡在广州，课程1在80分以上，且课程9在9分以上的男同学的数量。(备注：该处做了修正，课程10数据为空，更改为课程9)
3. 比较广州和上海两地女生的平均体能测试成绩，哪个地区的更强些？
4. 学习成绩和体能测试成绩，两者的相关性是多少？（九门课的成绩分别与体能成绩计算相关性）

**实验实施**：

（在此详述平台，技术栈，思路，处理逻辑等等）

平台：python

用到的函数包：cmath，pandas

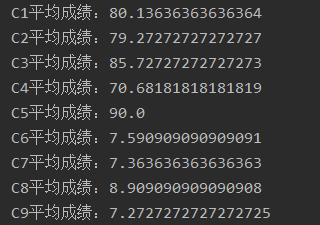
用来camth开方，pandas用来读取数据，合成两个数据表

数据：文件123.txt和123.xlsx是老师提供的两个数据，data.xlsx是两个数据合成后经过处理的数据，2.py是实验二的代码

**运行结果**：

（运行结果、现象的说明与截图）

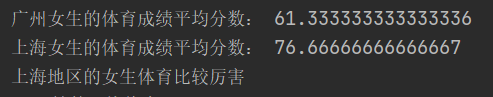
1.学生中家乡在Beijing的所有课程的平均成绩。



2.学生中家乡在广州，课程1在80分以上，且课程9在9分以上的男同学的数量。(备注：该处做了修正，课程10数据为空，更改为课程9)

ff47da1310434dcc1be4a382cf58878

1. 比较广州和上海两地女生的平均体能测试成绩，哪个地区的更强些？



分数分别按照100分，80分，60分和0分算

1. 学习成绩和体能测试成绩，两者的相关性是多少？（九门课的成绩分别与体能成绩计算相关性）





**实验总结**：

（遇到的问题，有什么感想和收货等等）

学会了如何用pandas去处理excel和csv的数据，怎么去合成两个类型不同的数据表，去查重和填充，去百度python如何使用，在这过程中学到了很多有关于机器学习的知识，在组员之间的相互配合下，我们终于完成了这个实验。