**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 12 月 1日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学 院** | 计算机科学与网络工程学院 | **年级/专业/班** | 软件183 | **姓名** | 谢先新 | **学号** | 1806300020 |
| **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\*\*\*)

**一、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

**二、实验设备与平台**

1. 实验设备：计算机；

2. 平台：Windows 10操作系统

**三、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）

1. 文件说明：数据源1：data.xlsx

数据源2：txt.txt

集成、清洗统一后的数据：dataall.csv

1. 实验具体平台：devc++,win10操作系统，c++语言
2. 函数说明：1：void readdatabase(string file) {}读取数据库函数

2：int stringturn(string s) {}将string 类型转为int类型函数

3：void readtxt(string file) {}读取txt文件函数

4：void compose() {}合成数据库与txt函数

5：void read(){}将数据读取到内存函数

6：void show(){}输出内存中数据函数

7：void solve1(){}求第一问问题函数

8：void solve2(){}求第二问问题函数

9：void solve3(){}求第三问问题函数

10：void solve4(){}求第四问问题函数

11：float mean()计算均值函数

12：float correlation()计算相关性函数

13：void clean()回收内存函数

14：int main()主函数

1. 调用的库函数：sqrt()函数
2. 技术栈与思路逻辑：

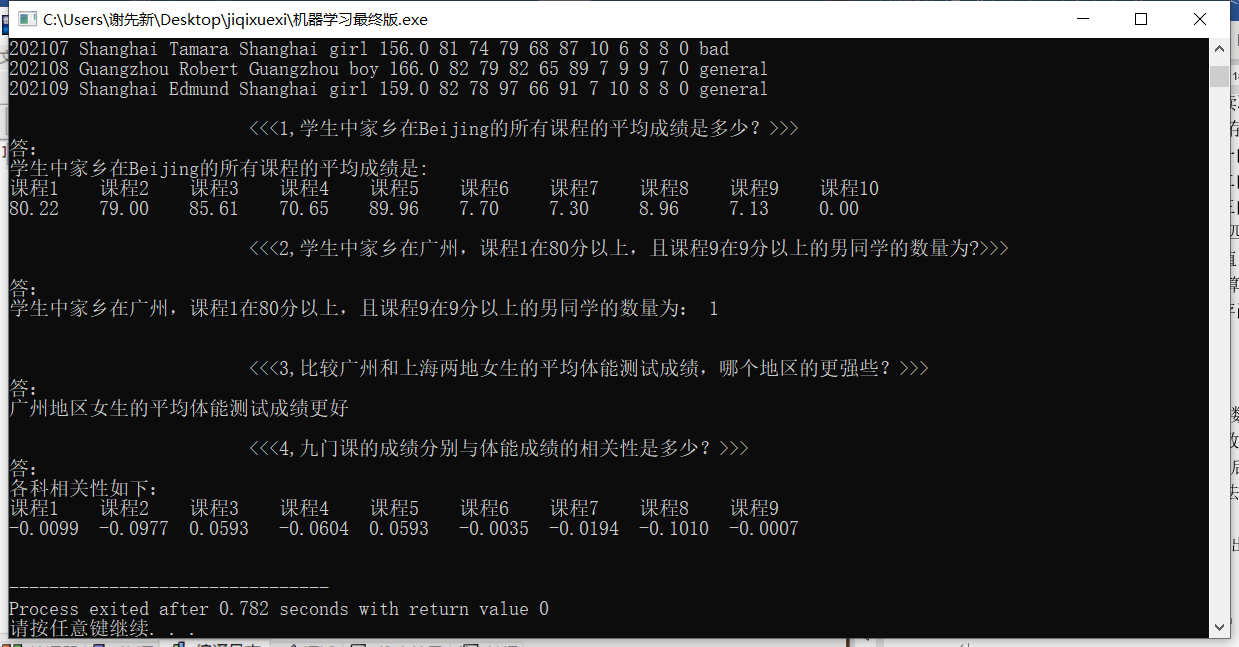
1：将老师给的两个数据源通过数据库连接后，通过创建数组来存 储数据，并通过数组来对数据进行操作。

2：通过1合成表csv后，又通过vector容器来对数据进行操作，在一定的数学公式和算法逻辑基础上，设计代码来解决实验一中给出的四个问题。

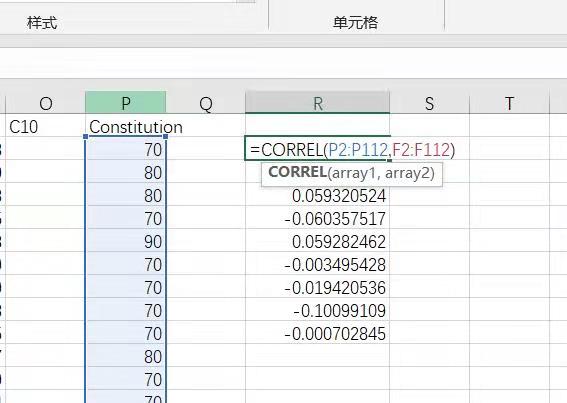
3：为了防止内存的溢出，最后还会对内存数据进行一个回收处理

**运行结果**：

（运行结果、现象的说明与截图）



从实验结果截图可以看到，我对合成的数据进行了一个输出，由于老师没有明确指明删除重复数据时是以哪一个参数为指标，所以我的合成表输出结果可能会与其他同学的个别数据会有所不同，但是总的来说，这并不影响相关性的结果。为了确保我的实验结果的正确性，我在合成数据的表中，利用公式已经算出了各科相关性，具体如下：



与我输出的实验结果进行比较可知，我输出的实验结果是正确的。

**实验总结**：

（遇到的问题，有什么感想和收货等等）

1. 遇到的问题

本次实验过程中遇到的主要问题如下：

1. 首先第一次接触机器学习知识，加上本身英语不过关，老师教学过程中是使用英文教学的，所以学习过程中对我来说是十分困难的，因此，由于机器学习知识的不牢固，在实验开始阶段是毫无头绪的
2. 对于读取文件中的数据这块，是之前数据结构中所学习的文件io流方面的知识，但是那时候由于老师没有细讲，所以在文件io流这一块是又自学了一段时间才比较掌握。
3. 会读取文件数据后，还要将数据进行合并，清洗和同一，在这过程中还需要连接数据库，这些对我来说，都基本算是没掌握的知识，所以这一块也是比较费时
4. 在代码编写过程中也是遇到各种麻烦，主要的问题是还是在数据的合并和读取中，但是搞定了这一块之后，最后用vector容器对数据进行操作来解决实验四个问题还是比较简单的，都是按照公式通过代码实现就行了。
5. 感想和收获

我终于明白什么是置之死地而后生了，通过这个实验，我终于对以前落下的知识有了一定的掌握，那就是对文件数据进行读取，合并，清洗等方面的知识。在这过程中，还需要很强的逻辑思维，另外还要足够细心，就比如由于自己是使用dev这个轻装版软件来编写的，在编译过程中是完全没有提示功能的，只能在运行时才能知道自己哪里出错，由于自己将一个参数赋值错了，导致结果一直不对，而编译和运行也是无法查到出错的，因此，编写代码时一定要细心、细心再细心

另外，通过这次实验，我也对机器学习有了一定了解，也重新学习了一下数学方面的知识。不管怎样，尽管这次实验的路程是十分坎坷的，但是收获还是很大的，老师也是想通过实验来培养我们学习知识的能力，通过实验来更好了解什么是机器学习。

由于我学习的语言是c++，而对于实验二，是要用到MFC等方面的知识，而且自己对那方面的知识基本是毫无头绪的，加上自己其他课程的任务量远比这门课程的多，接下来还要相继准备各门考试，所以在综合评估自己的实力还有课程情况来看，我就不再进行做实验二、三了，也希望老师能体谅，在此由衷感激！