# 学生成绩预测竞赛实验计划

班级：计科一班

队名：菜鸡组合

成员：汪东启、滕云鹏

**一、题目背景**  
 学生在校园中会产生各种行为数据，希望可以通过大数据分析的方法通过对学生的日常行为进行分析，挖掘用户的作息规律、兴趣爱好等，精确地预测学生之间的相对排名。通过对这些日常行为的建模来预测学生的学业成绩，从而可以实现提前预警学生的异常情况，并进行适当的干预，因而对学生的培养、管理工作将会起到极其重要的作用。  
**二、任务要求** 本次竞赛中，将从某高校的某个学院随机抽取一定比例学生，提供这些学生在三个学期的图书馆进出记录、一卡通消费记录、图书馆借阅记录、以及综合成绩的相对排名。这一部分数据将作为训练数据。我们从另外的某学院随机抽取一定比例的学生，然后提供他们在三个学期的图书馆进出记录、一卡通消费记录、图书借阅记录、以及前两个学期的成绩排名。第三学期的成绩排名作为预测目标。  
**三、数据格式及文件组成** 训练\成绩.csv。训练集的成绩文件，包含学期、学号、以及相对排名

训练\借书.csv。训练集的图书借阅信息，包含学期、学号、书号、日期

训练\图书门禁.csv。 训练集的图书门禁进入，包含学期、学号、日期、时间

训练\消费.csv。训练集的消费数据，包含学期、学号、地点、日期、时间、金额

测试\成绩.txt。 测试集的成绩文件。字段同上

测试\借书.txt。 测试集的图书借阅信息。字段同上

测试\图书门禁.txt。 测试集的图书门禁进入。字段同上

测试\消费.txt。 测试集的消费数据。字段同上  
**四、评估函数** 算法通过衡量预测排名和实际排名的Spearman相关性，为[0,1]之间的值，值越大，表示越相关，排名的预测就越准确。若要考虑n个学生的排名，学生i的预测排名为pi，而实际的排名为ri，di = pi - ri，那么

  
**五、实验步骤** **1）、数据读取**  
 读取提供的所有相关文件，读入图书的类别以及每个学生的图书馆门禁、成绩、结束情况、消费情况等信息，通过读取相应的信息完成后续的处理。

**2）、数据预处理** 对读取的数据进行预处理，对数据中的错误部分进行修改操作，对数据中的这些信息进行删除；数据中还有部分数据缺失，将这部分数据全部补为0。暂定将数据最终格式化为：  
x\_data = [学期 学号 图书馆门禁次数 食堂总消费...(各类消费的情况) 各类书的借阅次数 两个学期的排名]  
y\_data = [第三学期的排名]  
 **3）、训练模型** 暂定使用scikit-learn工具包中的随机森林算法解决问题。将之前格式化好的数据放入随机森林算法中进行训练。  
 **4）、数据测试并对比结果** 对测试集进行相同的预处理，处理完成后用对训练好的模型训练集进行预测，将预测结果用评估函数进行评估，记录准确率。  
 **5）、调节参数，提高准确度** 对前面的步骤中的部分参数进行调节，重复测试，尝试不同的方法提高预测的精确度。  
**六、涉及知识**

数据预处理

csv文件格式及读取

随机森林算法原理及应用  
**七、题目链接** http://www.dcjingsai.com/common/cmpt/%E5%AD%A6%E7%94%9F%E6%88%90%E7%BB%A9%E6%8E%92%E5%90%8D%E9%A2%84%E6%B5%8B\_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html