2019-2020学年第2学期

《大数据应用开发》期末项目报告

**项 目： 幸福指数分析**

**学 号： 17561140112**

**姓 名： 张峻溥**

**班 级： 17.1**

**学 院： 计算机与信息技术学院**

2020年6月

一、项目背景

|  |
| --- |
| **项目背景**  幸福是指人类在自身满足和安全的基础上主观地产生的一系列愉悦、舒适的情感。  满足人们对幸福追求的方式，主要表现在前现代时期哲学家的思考和人们对个人生活方式的选择上，而在现代时期则表现为投入到社会运动和制度建设中，体现了一种历史的进步。现代化给人类带来的重要成就之一，无疑是物质生活条件的不断改善和生活质量的不断提高。然而，现代化也是一个充满矛盾的过程。与客观幸福感的提高相比，主观幸福感并没有相应的增加，这无疑构成了现代化的困境。  如果我们能发现影响幸福的共性，生活将会变得更有趣。如果能找到影响幸福感的政策因素，就可以优化资源配置，提高国民幸福感。目前社会科学研究的重点是变量的解释能力和未来政策的执行情况，主要采用线性回归和逻辑回归的方法。并对政府公共服务、宏观经济环境、税负等宏观因素进行了一系列的推测和发现。  **项目要求**  本项目尝试幸福预测这一经典课题，希望在现有的社会科学研究之外尝试其他维度的算法，结合各学科的优势，探索潜在的影响因素，找到更容易解释和理解的相关性。  **数据集介绍**  该数据集来自中国人民大学中国调查与数据中心主持之《中国综合社会调查（CGSS）》项目。赛题感谢此机构及其人员提供数据协助。中国综合社会调查为多阶分层抽样的截面面访调查。主要字段包括，采访地点-省/自治区/直辖市编码、性别、出生日期-年、民族、宗教信仰-不信仰宗教、最高教育程度、个人去年全年的总收入、政治面貌、住房的套内建筑面积、房子的产权属于、身高、体重、身体健康状况、户口登记状况、户口登记地等字段与取值含义，对照表可参考附件happiness\_index.xlsx |

二、分析过程

|  |
| --- |
| **数据的预处理**   1. 读入数据，包括训练集和测试集，输出信息以确认。        1. 处理缺失/无意义的值   因本数据来自问卷调查，存在较多的缺失值，因此处理项较多。  1.查看缺失值的列及比率     1. 对部分缺失数据进行填充，并去除部分不重要的特征以用来消除噪点，基本方法是将值为“未知、无法回答”的值替换为平均值或众数值，同时删除空列。     **数据可视化**   1. 针对训练集的幸福指数分布进行统计与可视化       可以发现众数为4且结果不均匀，符合实际情况，表明当今社会大多数人是比较幸福的。   1. 可视化学历与幸福感的关系     可以看出认为自己幸福的人群中，高学历占比较低学历更高一些，基本符合社会现状。  **模型建立**   1. 输出相关性矩阵和相关性系数        1. 根据相关系数最强的几个变量作为模型输入      1. 创建模型并训练     **预测效果** |
|  |

三、心得体会（不少于400字）

|  |
| --- |
|  |
|  |

四、自我评分

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学号 | 姓名 | 数据预处理  20分 | 模型建立  30分 | 参数调优  10分 | 预测效果  10分 | 数据可视化  10分 | 项目报告  20分 | 总分  100分 |
| 17561140112 | 张峻溥 | 18 | 25 | 7 | 7 | 7 | 18 | 82 |