1．课程设计的目的

《统计过程与数据挖掘课程设计》的目的是为了让学生在学习《统计过程与数据挖掘》课程的基础上，进一步深入理解数据挖掘的基本理论,并将理论知识和实践结合起来，学生可以初步理解数据挖掘常见算法原理，可以通过 Python 编程实现数据挖掘项目，经过学习后续专业课程最终可以胜任数据分析工程师、算法工程师等相关岗位工作。

2．课程设计题目和要求

题目：基于算法的##研究（**与##处根据选择的任务标题填写）**

要求：独立按时完成实验，严格遵循具体任务内容要求，实验结果表述规范，报告字体与图表格式清晰，提交完整实验报告。

注：

1. 未注明“选择其 x 即可/选择若干即可”的均指全部。
2. 任务 1-4 的选择人数分别均不得超过班级人数的 30%。任务 5 的选择人数不得超过班级人数的 10%。

任务 1：基于回归分析的卫生选择行为研究

1．了解常见回归模型（含一元线性回归、逻辑回归、多项式回归等）的原理；

2．了解 CFPS 数据；

3．了解并运用常见的数据预处理与标准化处理方法；

4．掌握使用 sklearn 进行回归模型的训练、预测、评估(含 MSE、RMSE、MAE、r2_score 等，选择合适的评估指标若干)；

5.根据实验结果分析人群的卫生选择行为。

任务 2：社会科学标题的学科分类研究

1. 了解经典分类算法（含决策树、朴素贝叶斯、支持向量机等）的基本原理；
2. 理解并运用常见的数据集划分方法；

3. 了解并运用简单的文本向量化方法（含 TF-IDF 等）；
4. 掌握使用 sklearn 进行经典分类算法的训练、预测并评估模型性能；
5. 根据分类结果（结合混淆矩阵）分析不同学科间的关系。

任务 3：社会科学标题的学科聚类研究

1. 了解经典聚类算法（含 K-means、层次聚类、DBSCAN 等）的基本原理；
2. 理解并运用常见的数据集划分方法；
3. 了解并运用简单的文本向量化方法（含 TF-IDF 等）；
4. 了解并运用常见的距离度量方法；
5. 掌握使用 sklearn 进行经典聚类算法的训练、预测并评估聚类质量；
6. 根据聚类结果分析不同学科间关系。

任务 4：基于深度学习模型的社会科学标题学科分类研究（拓展推荐*）

1. 了解经典深度学习算法（Text-CNN、RNN、LSTM、BERT，选择其一即可）的基本原理；
2. 理解并运用常见的数据集划分方法；
3. 阅读并初步理解官方提供的经典深度学习算法代码；
4. 调试运用官方提供的代码进行训练、测试并评估不同超参数下的分类性能。
5. 根据分类结果（结合混淆矩阵）分析不同学科间的关系。

任务 5：其他有价值的研究

（选择该项需提供详细的研究方案并经老师审核通过）

1. 了解该领域常用的算法模型以及相关研究（要求阅读 15 篇 CSCD、CSSCI 或 8 篇 SCI、SSCI 研究型文献）；
2. 理解并运用该领域常见的数据处理方法；
3. 能够初步了解并调试相关代码进行模型的训练、测试、评估；
4. 根据实验结果分析并得出结论。

3．课程设计任务及工作量的要求〔包括课程设计计算说明书、图纸、实物样品等要求〕

（1）任务：

① 学习数据挖掘的基本理论，掌握初步的数据挖掘方法；

② 学会查阅技术资料和手册；

③ 掌握 python 开发工具的使用方法；

④ 提高综合运用所学的理论知识来分析问题和解决问题的能力；

⑤ 撰写规范的课程设计报告，培养严谨的作风和科学的态度。

（2）工作量要求：

① 阅读了解相关模型的经典研究文献并撰写简单的文献综述（不低于 10 篇 CSCD、CSSCI 期刊刊载的研究论文，或不少于 5 篇的 SCI、SSCI 期刊收录论文或综述）；

② 掌握常见的数据预处理方法并对实验中的数据处理、数据集划分流程给出充分说明与解释；

③ 科学合理设计不同模型的对比试验；

④ 根据实验结果不断调整参数优化模型性能；

⑤ 结合文献调研与实验结果，就未来进一步研究给出设想方案；

（3）验收：

学生根据课程设计的要求完成任务后，保存模型训练的结果和可视化的图表，并向指导教师请求验收。对达到设计要求的，教师将对其进行综合应用能力和操作能力的考核，然后给出实际操作能力分数。对未达到设计要求的学生，指导老师提出改进意见，学生完成后再提交验收。

验收从选题难度系数、任务要求完成度、代码的规范性与完整性等方面进行评估。

（4）课程设计报告（含研究创新能力评估）：

课程设计报告应包括如下几个部分：课程设计目的和意义，课程设计任务及要求，算法设计与分析，源程序，结果及分析，课程设计总结，参考资料等。

参考资料的著录需参照 GBT7714 国家标准。

课程设计报告的评估从报告撰写的规范性、研究型报告的综合写作水平、数据处理与实验方法的严谨性、研究创新能力等方面进行评估。

4．主要参考文献

[1]李航编著，《统计学习方法》，清华大学出版社，2012 年 3 月.

[2]周志华编著，《机器学习》，清华大学出版社，2016 年 1 月.