【数据分析与挖掘】

【Data Analysis and Mining】

一、基本信息

课程代码：【 2050544 】

课程学分：【3】

面向专业：【**软件**工程】

课程性质：【专业限选课】

开课院系：信息技术学院 软件工程系

使用教材：

教材【Python 数据分析从入门到精通，明日科技，清华大学出版社，2021 年 6 月】

参考书目【Python 数据分析与挖掘实战（第 2 版），张良均 谭立云 刘名军 江建明 机械工业出版社，2019 年 4 月】

课程网站网址：https://edu.csdn.net/

先修课程：【Java 程序设计(双语)2050010（3）】

【数据库原理 2050217（3）】

二、课程简介

通过本课程的学习，使学生学会使用 Python 语言进行数据探索、数据预处理、分类与预测、聚类分析、时序预测、关联规则挖掘、智能推荐、偏差检测等操作，并完成大量数据挖掘工程案例，将理论与实践相结合，让学生熟练掌握使用 Python 语言对样本数据进行处理、挖掘建模，为将来从事数据分析研究、工作奠定基础。

本课程将采用理论与实践相结合的教学方法。在理论上，通过项目引入概念、原理和方法。在实践上，充分地利用现有的硬件资源，发挥学生主观能动性，指导学生使用 Python 语言进行数据探索、数据预处理、分类与预测、聚类分析、时序预测、关联规则挖掘、智能推荐、偏差检测等。同时结合若干综合案例，引导学生将所学知识与企业需求相结合，将知识活学活用。

三、选课建议

本课程适合计算机科学与技术、物联网工程、数据科学与大数据技术、信息安全、网络工程、软件工程专业（本科）三年级学生。建议在第六学期开设。

四、课程与专业毕业要求的关联性

软件工程专业毕业要求	关联
L01：工程知识：能够将数学、自然科学、工程基础和专业知识用于解决复杂工程问题	
L02：问题分析：能够应用数学、自然科学和工程科学的基本原理，识别、表达、并通过文献研究分析复杂工程问题，以获得有效结论	
L03：设计/开发解决方案：能够设计针对复杂工程问题的解决方案，设计满足特定需求的系统、单元（部件）或工艺流程，并能够在设计环节中体现创新意识，考虑社会、健康、安全、法律、文化以及环境等因素	●
L04：研究：能够基于科学原理并采用科学方法对复杂工程问题进行研究，包括设计实验、分析与解释数据、并通过信息综合得到合理有效的结论	●
L05：使用现代工具：能够针对复杂工程问题，开发、选择与使用恰当的技术、资源、现代工程工具和信息技术工具，包括对复杂工程问题的预测与模拟，并能够理解其局限性	●
L06：工程与社会：能够基于工程相关背景知识进行合理分析，评价专业工程实践和复杂工程问题解决方案对社会、健康、安全、法律以及文化的影响，并理解应承担的责任	
L07：环境和可持续发展：能够理解和评价针对复杂工程问题的专业工程实践对环境、社会可持续发展的影响	
L08：职业规范：具有人文社会科学素养、社会责任感，能够在工程实践中理解并遵守工程职业道德和规范，履行责任	
L09：个人和团队：能够在多学科背景下的团队中承担个体、团队成员以及负责人的角色	
L010：沟通：能够就复杂工程问题与业界同行及社会公众进行有效沟通和交流，包括撰写报告和设计文稿、陈述发言、清晰表达或回应指令。并具备一定的国际视野，能够在跨文化背景下进行沟通和交流	
L011：项目管理：理解并掌握工程管理原理与经济决策方法，并能在多学科环境中应用	
L012：终身学习：具有自主学习和终身学习的意识，有不断学习和适应发展的能力	

五、课程目标/课程预期学习成果

序号	课程预期学习成果	课程目标（细化的预期学习成果）	教与学方式	评价方式

1	L032 能针对需求分析独立进行算法设计和程序实现，并能测试验证算法与程序的正确性	1. 能够熟练掌握 Python 开发环境的搭建和核心语法	讲授、练习	实验报告 课堂展示
		2. 掌握 Python 进行数据分析的函数包的使用方法	讲授、练习	
		3. 掌握 Python 进行数据可视化的函数包的使用方法	讲授、练习	
		4. 掌握 Python 进行机器学习的函数包的使用方法	讲授、练习	
2	**L041 能够基于科学原理，结合软件行业，通过文献研究等相关方法，调研和分析软件系统设计问题**	能够以团队的形式帮助团队中其他学习有困难的同学，帮助他们战胜学习上的困难，培养他们学习兴趣和开发能力	自主学习、团队学习	实验报告
3	L052 能够针对具体复杂软件工程的特点和需求，选择合适的开发环境或技术工具进行设计开发，或使用模拟软件进行模拟	能够利用课后的扩展阅读，了解行业的前沿知识技术，并能通过团队的力量进行协作学习、共同探究了解到的前沿知识技术，并能在软件或软件的某一模块中运用	课后阅读、自主学习、团队讨论、协作开发	大作业 课堂展示

六、课程内容

- 掌握什么是数据分析
- 了解数据分析的重要性
- 了解数据分析的基本流程
- 熟练使用数据分析常用工具

理论学时 2 学时

第二单元 搭建 Python 数据分析环境

- 了解 Python 概述
- 掌握搭建 Python 开发环境
- 掌握集成开发环境 PyCharm
- 掌握数据分析标准环境 Anaconda

-掌握 Python 核心语法和数据结构

实验课时 4 学时，理论学时 4 学时

第三单元 Pandas 统计分析

- 了解 Pandas 、 Series 对象 、 DataFrame 对象
- 熟练导入外部数据 、对数据进行管理操作，以及数据清洗，并进行以下操作
- 索引设置
- 数据排序与排名
- 数据计算
- 数据格式化
- 数据分组统计
- 数据移位
- 数据转换
- 数据合并
- 数据导出
- 日期数据处理
- 时间序列
- 能够运用以上知识进行综合应用

实验课时 4 学时，理论学时 2 学时

第四单元 可视化数据分析图表

- 了解 Matplotlib 、Seaborn、第三方可视化数据分析图表 Pyecharts 的作用
- 了解图表的基本组成，掌握如何选择适合的图表类型
- 掌握图表的常用设置 ，以及常用图表的绘制
- 能够根据不同的数据需求，选择不同的可视化数据分析图表，并进行综合应用

实验课时 4 学时，理论学时 4 学时

第五单元 图解数组计算模块 NumPy

- 掌握 NumPy 的基本操作
- 熟练使用 NumPy 常用统计分析函数
- 对需求进行分析，并能综合应用

理论学时 4 学时

第六单元 数据统计分析案例

- 对比分析
- 同比、定比和环比分析
- 贡献度分析（帕累托法则）
- 差异化分析
- 相关性分析
- 时间序列分析

实验学时 4 学时，理论学时 6 学时

第七单元 机器学习库 Scikit-Learn

- Scikit-Learn 简介
- 安装 Scikit-Learn
- 线性模型
- 支持向量机
- 聚类

实验学时 4 学时，理论学时 2 学时

第八单元 项目案例分析

掌握 4 个典型案例的分析，包括 注册用户分析、电商销售数据分析与预测、 二手房房价分析与预测和 客户价值分析

实验学时 4 学时

七、课内实验名称及基本要求

列出课程实验的名称、学时数、实验类型（演示型、验证型、设计型、综合型）及每个实验的内容简述。

序号	实验名称	主要内容	实验时数	实验类型	备注
1	Python 编程	使用 Python 编写应用程序	6	设计型	Anaconda 3.x 以上版本，MySql5.5 以上
2	Python 函数包的使用	使用 Numpy, Pandas,Matplotlib,Pyecharts,Seaborn,Scikit-Learn 函数包进行数据分析与挖掘	6	设计型	同上
3	综合应用	综合利用 Python 语言编写不同应用场景下的数据分析与挖掘的小型项目	12	设计型	同上

八、评价方式与成绩

总评构成（1+X）	评价方式	占比
1	大作业	60%
X2	实验报告	25%
X3	课堂表现	15%

撰写人：刘俊　　系主任审核签名：朱丽娟

审核时间：2022 年 2 月 10 日