课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	课程名称Course Title	**数据科学实践 / Data Science Practice**
2.	授课院系 Originating Department	统计与数据科学系 Department of Statistic and Data Science
3.	课程编号 Course Code	STA326
4.	课程学分Credit Value	3
5.	课程类别 Course Type	专业核心课 Major Core Courses
6.	授课学期 Semester	春季 Spring
7.	授课语言 Teaching Language	中英双语 English & Chinese
8.	授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）	魏鸿鑫 Wei Hongxin 统计与数据科学系Department of Statistics and Data Science weihx@sustech.edu.cn
9.	实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact	待公布 To be announced
10.	选课人数限额(可不填) Maximum Enrolment（Optional）	

11. 授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
学时数	32		32		64

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Credit Hours					

12.	**先修课程、其它学习要求 Pre-requisites or Other Academic Requirements**	STA303 人工智能 B Artificial Intelligence B
13.	**后续课程、其它学习规划 Courses for which this course is a pre-requisite**	
14.	**其它要求修读本课程的学系 Cross-listing Dept.**	

教学大纲及教学日历SYLLABUS

15. 教学目标Course Objectives

1. 理解数据科学实践的基本概念：课程旨在帮助学生们理解数据科学实践的核心概念，包括数据收集与处理、特征工程、机器学习建模、模型评估和部署等方面。
2. 掌握数据科学工具和技术：课程将提供学生们学习和使用常见的数据科学工具和技术的机会，如 Python 编程语言、数据处理库（如 Pandas 和 NumPy）、机器学习库（如 Scikit-learn 和 TensorFlow）以及数据可视化工具（如 Matplotlib 和 Tableau）等。
3. 培养实际问题解决能力：通过实践项目和案例，学生们将学习如何应用数据科学方法解决真实世界的问题。他们将学会提出问题、设计和执行实验、分析和解释结果，并提出有效的解决方案。
4. 团队合作与沟通能力：课程将鼓励学生们进行团队合作项目，培养他们的团队合作和沟通能力。他们将学会与团队成员合作，共同解决问题，并有效地传达他们的分析和结论。

1. Understand the basic concepts of data science practice: The course aims to assist students in comprehending the core concepts of data science practice, encompassing aspects such as data collection and processing, feature engineering, machine learning modeling, model evaluation, and deployment.
2. Master data science tools and techniques: The course will provide students with opportunities to learn and utilize common data science tools and techniques, including the Python programming language, data processing libraries (such as Pandas and NumPy), machine learning libraries (such as Scikit-learn and TensorFlow), as well as data visualization tools (such as Matplotlib and Tableau).
3. Cultivate practical problem-solving skills: Through practical projects and case studies, students will learn how to apply data science methods to solve real-world problems. They will acquire the ability to formulate clear problem statements, design and execute appropriate experiments, analyze and interpret results, and propose effective solutions.
4. Foster teamwork and communication skills: The course will encourage students to engage in team-based projects to cultivate their teamwork and communication skills. They will learn to collaborate effectively with team members, collectively address problems, and communicate their analyses and conclusions efficiently.

16. 预达学习成果Learning Outcomes

通过参加数据科学实践课程，学生将理解数据科学实践的基本概念，掌握数据科学工具和技术，培养实际问题解决能力，以及发展团队合作和沟通能力。他们将能够应用数据科学方法解决真实世界的问题，并具备使用常见数据科学工具和技术的能力。这门课程将为学生在数据科学领域的职业发展提供坚实的基础。

By participating in the Data Science Practicum course, students will gain an understanding of the fundamental concepts of data science practice, master data science tools and techniques, develop practical problem-solving skills, and enhance teamwork and communication abilities. They will be able to apply data science methods to solve real-world problems and possess the proficiency to use common data science tools and technologies. This course will provide a solid foundation for students' career development in the field of data science.

17. 课程内容及教学日历
（如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

理论课程

1. 数据科学简介 (2 学时)
   - 数据科学概述及其在不同领域的应用
   - 数据科学项目中的道德考量和数据隐私
   - 数据科学生命周期和最佳实践介绍

2. 数据收集与预处理 (4 学时)
   - 从不同来源（如数据库、API、网络抓取）获取数据的技术
   - 数据清洗、数据整理和处理缺失值
   - 探索性数据分析和数据可视化技术

3. 特征工程与选择 (2 学时)
   - 从原始数据中转换和创建有意义特征的方法
   - 特征缩放、归一化和编码方法
   - 特征选择和降维技术

4. 机器学习建模 (4 学时)
   - 监督学习、无监督学习和半监督学习算法的介绍
   - 常见的机器学习算法，如线性回归、决策树、支持向量机、聚类等
   - 模型评估和选择，包括交叉验证和性能指标

5. 模型部署与优化 (2 学时)
   - 模型部署的基本概念和常用方法
   - 模型性能优化和调整超参数的技巧
   - 实施预测和实时推理的方法

6. 金融行业应用案例 (2 学时)
   - 金融数据分析与风险管理案例
   - 投资组合优化与资产定价案例
   - 金融欺诈检测与反洗钱案例

7. 制造业应用案例 (2 学时)
   - 生产过程优化与质量控制案例
   - 故障预测与维修优化案例
   - 供应链可视化与供应商评估案例

8. 游戏行业应用案例 (2 学时)
   - 用户行为分析与游戏推荐系统案例
   - 游戏运营与用户留存优化案例
   - 虚拟经济与物品交易分析案例

9. 社交媒体数据科学应用案例 (2 学时)
   - 社交网络分析与影响力评估案例
   - 用户情感分析与舆情监测案例
   - 广告定向投放与推荐系统案例

10. 零售行业应用案例 (2 学时)
   - 市场篮子分析与销售趋势预测案例
   - 个性化推荐与客户细分案例
   - 供应链优化与库存管理案例

11. 未来方向讨论班 (2 学时)
   - 新兴应用领域讨论
   - 技术趋势讨论
   - 数据伦理和隐私保护
   - 挑战与机遇

12. 课程项目报告（6 学时）
   - 项目背景和目标介绍
   - 方法过程描述
   - 结论描述及问题总结

实验课程

第 1 部分：数据科学工具练习（8 学时）
   - Linux bash 及 Jupyter 工具学习
   - Python 编程基础与数据分析库
   - 团队协作工具 Git 学习

第 2 部分：数据预处理与探索性数据分析实践（8 学时）
   - 实际数据集的导入与清洗
   - 数据转换与规范化
   - 数据可视化实践
   - 数据描述性统计分析实践

第 3 部分：数据挖掘与机器学习实践（8 学时）
   - 利用实际数据集进行数据挖掘

- 监督学习算法实践
- 无监督学习算法实践
- 模型评估与选择

第 4 部分：课程项目实践 （8 学时）

- 项目规划和管理
- 团队合作和沟通技巧
- 问题解决和决策能力

Lectures

1. Introduction to Data Science (2 hours)
    - Overview of data science and its applications in different domains
    - Ethical considerations and data privacy in data science projects
    - Introduction to the data science lifecycle and best practices
2. Data Collection and Preprocessing (4 hours)
    - Techniques for data acquisition from various sources such as databases, APIs, web scraping
    - Data cleaning, data wrangling, and handling missing values
    - Exploratory data analysis and data visualization techniques
3. Feature Engineering and Selection (2 hours)
    - Methods for transforming and creating meaningful features from raw data
    - Feature scaling, normalization, and encoding techniques
    - Feature selection and dimensionality reduction techniques
4. Machine Learning Modeling (4 hours)
    - Introduction to supervised, unsupervised, and semi-supervised learning algorithms
    - Common machine learning algorithms such as linear regression, decision trees, support vector machines, clustering, etc.
    - Model evaluation and selection, including cross-validation and performance metrics
5. Model Deployment and Optimization (2 hours)
    - Basics of model deployment and common methods
    - Techniques for model performance optimization and hyperparameter tuning
    - Implementation of predictions and real-time inference
6. Application cases in the financial industry (2 hours)
    - Financial data analysis and risk management cases
    - Portfolio optimization and asset pricing cases
    - Financial fraud detection and anti-money laundering cases
7. Application cases in the manufacturing industry (2 hours)
    - Production process optimization and quality control cases
    - Fault prediction and maintenance optimization cases
    - Supply chain visualization and supplier evaluation cases
8. Application cases in the gaming industry (2 hours)

- User behavior analysis and game recommendation system cases
- Game operation and user retention optimization cases
- Virtual economy and item trading analysis cases

9. Data science application cases in social media (2 hours)
    - Social network analysis and influence assessment cases
    - User sentiment analysis and public opinion monitoring cases
    - Targeted advertising and recommendation system cases
10. Application cases in the retail industry (2 hours)
    - Market basket analysis and sales trend prediction cases
    - Personalized recommendation and customer segmentation cases
    - Supply chain optimization and inventory management cases
11. Discussion on future directions (2 hours)
    - Discussion on emerging application areas
    - Discussion on technology trends
    - Data ethics and privacy protection
    - Challenges and opportunities
12. Course Project Report (6 hours)
    - Introduction to Project Background and Objectives
    - Description of Methods and Procedures
    - Conclusion and Summary of Issues

Practical classes

Part 1: Data Science Tools Practice (8 class hours)

- Learning Linux bash and Jupyter tools
- Python programming basics and data analysis libraries
- Learning Git for team collaboration

Part 2: Data Preprocessing and Exploratory Data Analysis Practice (8 class hours)

- Importing and cleaning actual datasets
- Data transformation and normalization
- Data visualization practice
- Descriptive statistical analysis practice

Part 3: Data Mining and Machine Learning Practice (8 class hours)

- Data mining using actual datasets
- Supervised learning algorithm practice
- Unsupervised learning algorithm practice
- Model evaluation and selection

Part 4: Course Project Practice (8 class hours)

- Project Planning and Management
- Team Collaboration and Communication Skills
- Problem Solving and Decision Making Abilities

18. 教材及其它参考资料Textbook and Supplementary Readings

教材： Data Science Handbook, O'Reilly Media （中文版：数据科学手册）

参考资料：

Jake VanderPlas, Data Science Handbook, O'Reilly Media, Inc. 2016

Peter Harrington, 机器学习实战, 人民邮电出版社 2013

《数据科学家访谈录》（The Data Science Handbook） 人民邮电出版社

课程评估ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤Attendance		10		出勤率100%得满分，每缺席一次扣0.5分, 理论课和实验课共缺席超过 20 次扣满 10 分。
课堂表现 Class Performance				
小测验 Quiz				
课程项目Projects		40		分组完成一个数据科学相关的综合实践项目，最终上交代码和完整报告。

平时作业 Assignments		30		包括理论和实践两部分。理论作业要求学生掌握课程知识，实践作业要求学生运用所学知识完成编程任务。
期中考试 Mid-Term Test				
期末考试 Final Exam				
期末报告 Final Presentation		20		以小组为单位做项目展示及汇报。
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)				

20. 记分方式GRADING SYSTEM

√ A. 十三级等级制Letter Grading
☐ B. 二级记分制（通过/不通过） Pass/Fail Grading

课程审批REVIEW AND APPROVAL

21. 本课程设置已经过以下责任人/委员会审议通过
This Course has been approved by the following person or committee of authority