The Hong Kong University of Science and Technology (Guangzhou)

UG Course Syllabus Template

Introduction to Data Science and Analytics

DSAA 1001

3 Credits

Name: Wei Wang

Email: weiwcs@ust.hk

Name: Lei Li

Email: thorli@ust.hk

Name: Yongqi Zhang

Email: yongqizhang@hkust-gz.edu.cn

Course Description

Data science changes the way people process data in different areas. It has promoted the development of many subjects. This course introduces beginners to the whole lifecycle of data science problems and solutions. The course will help students comprehensively understand the basic knowledge of data science and use computer techniques to handle the real-life data science problems. Topics covered include data collection and processing, computer-oriented modelling techniques, relationship with other mathematics, and case studies.

Intended Learning Outcomes (ILOs)

By the end of this course, students should be able to:

1. Understand the whole lifecycle of data science problems and solutions.

2. Learn how to use computer-oriented modelling techniques to deal with data science problems.

3. Understand the relationship between data science and other mathematics.

4. Be able to handle some data science related cases.

Assessment and Grading

This course will be assessed using criterion-referencing and grades will not be assigned using a curve. Detailed rubrics for each assignment are provided below, outlining the criteria used for evaluation.

Assessments

Assessment Task	Contribution to Overall Course grade (%)	Due date
Module 1 Assignment	10%	TBD*
Module 2 Assignment	10%	TBD*
Final Exam	50%	TBD*
Project	20%	TBD*
Lab Attendance	5%	TBD*
Lecture Attendance	5%	TBD*

* Assessment marks for individual assessed tasks will be released within two weeks of the due date.

Mapping of Course ILOs to Assessment Tasks

Assessed Task	Mapped ILOs	Explanation
Module 1 Assignment	ILO2, ILO3, ILO4	This task assesses students' ability to understand knowledge and techniques related to probability distributions, estimation of statistics for large volume dataset, and popular estimation and inference methods (ILO2, ILO3, ILO4).
Module 2 Assignment	ILO1, ILO2, ILO4	This task assesses students' ability to understand and apply data management tasks with tools and algorithms. Data management is the first group of techniques and problems of data cycle (ILO1, ILO4) that requires its own problem modeling and tools (ILO2, ILO4).
Final Exam	ILO1, ILO2, ILO3, ILO4	This part also assesses understanding with respect to feature selection, decision tree classification model, logistic regression model and model evaluation such as precision and recall (ILO2, ILO3). Together with the first two modules, the exam will assess the student's foundational understanding of the data analysis pipeline (ILO1) and apply them to solve practical problem (ILO4).
Lab Project	ILO1, ILO2, ILO3, ILO4, ILO5	The project requires students to work through the whole data analysis process (ILO1) with their own codes (ILO2) on a real-life data analytic application (ILO3, ILO4)

Grading Rubrics

This course utilizes an absolute grading system as follows.

A: [100,85]; B: (85,70]; C: (70,55]; D: (55,40]; F: (40,0].

Subgrades (e.g., A-, B+, etc.) will be assigned for every 5 points within these ranges.

Final Grade Descriptors

Grades	Short Description	Elaboration on subject grading description
A	Excellent Performance	Demonstrates a comprehensive grasp of data science and analytics, expertise in problem-solving, and significant creativity in thinking. Exhibits a high capacity for scholarship and collaboration, going beyond core requirements to achieve learning goals.
B	Good Performance	Shows good knowledge and understanding of data science and analytics, competence in problem-solving, and the ability to analyze and evaluate issues. Displays high motivation to learn and the ability to work effectively with others.
C	Satisfactory Performance	Possesses adequate knowledge of data science and analytics, competence in dealing with familiar problems, and some capacity for analysis and critical thinking. Shows persistence and effort to achieve broadly defined learning goals.
D	Marginal Pass	Has threshold knowledge of data science and analytics, potential to achieve key professional skills, and the ability to make basic judgments. Benefits from the course and has the potential to develop in the discipline.
F	Fail	Demonstrates insufficient understanding of data science and analytics and lacks the necessary problem-solving skills. Shows limited ability to think critically or analytically and exhibits minimal effort towards achieving learning goals. Does not meet the threshold requirements for professional practice or development in the discipline.

1. Assignments

- Understanding of Concepts: Clear demonstration of key ideas and principles.
- Accuracy: Correct answers and calculations.
- Problem-Solving: Effective application of methods to solve problems.
- Clarity of Expression: Neat and organized presentation of answers.
- Implementation (Where applicable): Demonstrate clear evidence of the implementation.

2. Final Exam

- Understanding of Concepts: Clear demonstration of key ideas and principles.
- Accuracy: Correct answers and calculations.
- Problem-Solving: Effective application of methods to solve problems.
- Clarity of Expression: Neat and organized presentation of answers.

3. Lab Project

- Data Quality and Preprocessing: Problem identification and solution
- Model Design: Rational from model design and code originality
- Solution Efficiency: Data preprocessing time and model training time
- Model performance: Accuracy against a reasonable threshold

4. Attendance

- Participation: Engages actively during sessions.

Consistency: Attends regularly without excessive absences

Course AI Policy

1. Purpose and Scope This policy outlines the acceptable use of Generative AI (GenAI) tools in this course. It aims to ensure academic integrity while allowing students to leverage GenAI technologies to enhance their learning experience.

2. Allowed Uses Students are permitted to use GenAI tools for the following purposes:

- Idea Generation: To brainstorm ideas and explore different approaches to data science problems.
- Drafting and Refinement: To draft and refine reports, presentations, and other course-related materials.
- Code Assistance: To receive help with coding issues, debugging, and understanding complex algorithms.

3. Declaration Requirement Students must declare any use of GenAI tools in their submissions. This includes:

- Written Assignments: Clearly state in the assignment if GenAI tools were used. Describe the extent of their involvement (e.g., drafting text, generating code snippets).
- Presentations: Mention the use of GenAI tools in the acknowledgments or methodology section of presentations.

4. Academic Integrity The use of GenAI tools should not undermine academic integrity. Students are expected to:

- Understand and Learn: Ensure they understand the concepts and methods applied in their work, even if GenAI tools were used.
- Avoid Plagiarism: Do not use GenAI-generated content as their own without proper attribution.

5. Consequences of Misuse Failure to declare the use of GenAI tools or using them in a manner that breaches academic integrity may result in:

- Academic Penalties: Penalties as outlined in the course syllabus and university policies.
- Review: The work will be reviewed for compliance with this policy, and students may be required to explain their use of GenAI tools.

6. Support and Resources Students can seek guidance on the appropriate use of GenAI tools from course instructors and teaching assistants. Additional resources and tutorials on effective use will be provided.

7. Policy Review This policy will be reviewed periodically to ensure it remains relevant and effective. Updates will be communicated to students as necessary.

Communication and Feedback

Assessment marks for individual assessed tasks, including In-Class Quizzes and Lab Projects marks, will be communicated via Canvas within two weeks. Feedback on assignments will include comments on strengths and areas for improvement. Students who have further questions about the feedback including marks should consult the instructor within five working days after the feedback is received.

Resubmission Policy

Infinite resubmission before deadline, and no submission after deadline.

Required Texts and Materials

Lecture slides, software documentations, and research papers.

Academic Integrity

Students are expected to adhere to the university's academic integrity policy. Students are expected to uphold HKUST(GZ)'s Academic Honor Code and to maintain the highest standards of academic integrity. The University has zero tolerance of academic misconduct. Please refer to Regulations for Academic Integrity and Student Conduct for the University's definition of plagiarism and ways to avoid cheating and plagiarism.

Lecture	Topic	Goal	Lecturer	Lab	Assignment
L1	**Introduction + Course Info**	**Understand Data** **Statistic**	**Wei Wang**	**Lab 1 Environment**	**Assignment 1** **Out After L9** **Due Before L17**
L2	**Probabilistic Distributions**				
L3	**Unknown Distribution Estimation**				
L4	**Approximate Algorithm**				
L5	**Parameter Estimation I**			**Lab 2 Data Process**	
L6	**Parameter Estimation II**				
L7	**Latent Variable Model**				
L8	**Expectation Maximization**				
L9	**Module 1 Review**				
L10	**Data Collection**	**Manage Data** **Database**	**Lei Li**	**Lab 3 Data Collection**	**Assignment 2** **Out After L17**
L11	**Data Exploration + Visualization**				

L12	**Data Modeling and Storage I**				**Due Before L26**
L13	**Data Modeling and Storage II**				
L14	**Data Indexing**			**Lab 4 Data Management**	
L15	**Data Quality and Integration**				
L16	**Managing Big Data**				
L17	**Module 2 Review**				
L18	**Data Pre-Processing**	**Knowledge From Data** **Learning**	**Yongqi Zhang**	**Lab 5 Feature Engineering**	**Final Exam** **After L26**
L19	**Decision Tree I**				
L20	**Decision Tree II**				
L21	**Generalization Theorem**				
L22	**Ensemble I**			**Lab 6 Classification**	
L23	**Ensemble II**				
L24	**Clustering I**				
L25	**Clustering II**				
L26	**Module 3 Review**				