Lingnan University

Department of Computing and Decision Sciences

Course Syllabus

Course Title	:	Best Practices of Data Science
Course Code	:	CDS4001
Recommended Study Year	:	4
No. of Credits/Term	:	3
Mode of Tuition	:	Sectional Approach
Class Contact Hours	:	3 hours per week
Category in Major Prog.	:	Required
Prerequisite(s)	:	(a) CDS3004 Data Mining, and (b) CDS3001 Databases and Data Warehouses
Co-requisite	:	Nil
Exclusion	:	Nil
Exemption Requirement	:	Nil

Brief Course Description:

Data science is the study of where information comes from, what it represents and how it can be turned into a valuable resource in the creation of business and IT strategies. Mining large amounts of structured and unstructured data to identify patterns can help an organization rein in costs, increase efficiency, recognize new market opportunities and increase the organization's competitive advantage. This course elaborates on data science problems in science, social science, arts and business. The appropriate methods of delivering data science results to different domain users and the right processes for deploying results in information and/or intelligent systems are described. Practitioners in different fields, such as marketing and finance, will be invited to share their experience in applying the data science approach to solve their problems.

Aims:

This course aims to provide a crystal understanding of the whole data science process, combining theoretical concepts with real-life data science problems. Through the case studies in different domains, students will learn how to solve analytically complex data problems by using a blend of data inference, algorithm development, and technical skills step by step. The current/best practices of data science in Bioinformatics, Healthcare, Geographic Information System, Economics, Education, and e-Business will be discussed. Students will learn concepts, techniques, and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication. This course provides students hands-on experience with real-world data analysis. Students will apply critical thinking to analyze and solve data science problems.

Learning Outcomes (LOs):

On completion of this course, students will be able to:

1. Identify the fundamental principles and practical applications of data science in different domains (PLO1);

2. Describe and effectively apply the data science process and techniques to problem solving (PLO2);
3. Apply critical thinking skills to analyze data science problems and provide data support for decision making (PLO3);
4. Formulate the problems creatively and solve them using different algorithms and methods (PLO4);
5. Evaluate and validate the results obtained in a data science process (PLO7);
6. Interpret the results and deliver the findings (PLO8);
7. Implement programs based on various data structures and object-oriented programming (PLO9); and
8. Develop and deploy the database management system to solve some real-world problems (PLO9).

Indicative Contents:

Data Science
The cloud and related technologies
Trends in the emerging field of Data Science

What do data science people do?
Routines of Data Scienctists
R versus Python, and pros and cons of other common tools
Data science tools and technology

Data Science and Data Strategy
Achieving competitive advantage with data science
Sustaining competitive advantage with data science

Best Practices in the Data Science Process
- Problem understanding
  - Understand the business
  - Define the objectives
  - Build the objectives of the data analytic
- Data understanding
  - Data collection
  - Data imputation
  - Data type and structure
- Data preparation
  - Data cleansing
  - Data normalization
  - Data integration
  - Data transformation
- Data analytics and modeling
  - Statistical and other tests
  - Classification and Clustering
  - Prediction and Regression
  - Graph Analytics
  - Natural language processing
- Evaluation, Validation and Reporting
  - Evaluation metrics, such as the expected value approach
  - Profit curve, lift curve, and ROC curve
  - Data Reporting and Visualization Techniques

Case Studies of Data Science Application
Predictive Analytics for the Healthcare, Logistics, and Marketing domains
Machine Learning Approach for Corporate Knowledge Management
Real time Sales Nowcasting using Social Media and Sales Transaction Data
Other relevant cases in domains such as marketing and finance

Teaching Method:

There are different teaching and learning activities including lecture and laboratory sections. The concepts and principles of best practices of data science will be discussed in the lecture sections. The practical skills and techniques will be taught during the laboratory sections. Students are required to perform a group project to apply the concepts and principles covered in this course to critically analyze the given problem(s) and creatively formulate the solution(s). Students implement the data science solution(s) using relevant methods, and deliver a management report.

Assessment:

Class Attendance and Participation	5%
Assignments	20%
Midterm test	25%
Case Project	50%
Total	100%

Measurement of Learning Outcomes:

	Class Attendance and Participation	Assignments	Midterm test	Case Project
1. Recognize the fundamental principles and practical applications of data science in different domains (PLO1)	X		X	
2. Describe and effectively apply the data science process and techniques to problem solving (PLO2)	X	X	X	X
3. Apply critical thinking skills to analyze data science problems and provide relevant data for decision making (PLO3)		X	X	X
4. Formulate the problems creatively and solve them using different algorithms and methods (PLO4)		X	X	X
5. Evaluate and validate the result obtained in a data science process (PLO7)		X	X	X

	Class Attendance and Participation	Assignments	Midterm test	Case Project
6. Interpret the results and deliver the findings (PLO8)	X		X	X
7. Implement programs based on various data structures and object-oriented programming (PLO9)				X
8. Develop and deploy the database management system to solve some real-world problem (PLO9)		X		X

1. There are a number of classroom activities to evaluate if the students can recognize the fundamental principles and practical applications of data science in different domains. Students are expected to describe and effectively apply the data science process and techniques to problem solving. Students are required to interpret the results and deliver the findings (LO1-2. LO6, PLO1-2, and PLO8).
2. The assignments require students to describe and effectively apply the data science process and techniques to problem solving. Students need to apply critical thinking skills to analyze data science problems and provide data support for decision making. Students are required to formulate the problems creatively and solve them using different algorithms and methods, and evaluate and validate the results obtained in a data science process. Furthermore, students need to use a database management system to solve some real-world problems (LO2-5, LO8, PLO2-4, PLO7, and PLO9).
3. The Case Project requires students to describe and effectively apply the data science process and techniques to problem solving. Critical thinking skills will be used to solve data science problems and provide data support for decision making. Students are required to formulate the problems creatively and solve them using different algorithms and methods, and evaluate and validate the results obtained in a data science process. Students are expected to implement programs based on various data structures and object-oriented programming, and interpret the results and deliver the findings (LO2-8, PLO2-4, PLO7-9).
4. The midterm test will assess students' ability to recognize the fundamental principles and practical applications of data science in different domains. The data science process and techniques to problem solving will need to be described and applied effectively. Critical thinking skills will be applied. Problem formulation, evaluation, and result validation and interpretation will require students to apply principles and concepts learnt (LO1-6, PLO1-4, PLO7-8).

Required/Essential Readings:

1. Foster Provost, and Tom Fawcett, Data Science for Business: *What you need to know about data mining and data-analytic thinking*, O'Reilly, 2013.
2. Joel Grus, *Data Science from Scratch: First Principles with Python 1st Edition*, O'Reilly, 2015.

Recommended/Supplementary Readings:

1. Jure Leskovek, Anand Rajaraman, and Jeffrey Ullman, *Mining of Massive Datasets. v2.1*, Cambridge University Press, 2014.
2. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning*, *2nd Edition*, Springer, 2009.
3. John W. Foreman, *Data Smart: Using Data Science to Transform Information into Insight, 1st Edition,* 2013.
4. B. Baesens, *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications,* Wiley, 2014.
5. R. Bartlett, *A Practitioner's Guide to Business Analytics: Using Data Analysis Tools to Improve Your Organization's Decision Making and Strategy*, McGraw-Hill, 2013.
6. Mohammed J. Zaki, and Wagner Miera Jr., *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, 2014.
7. Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, *3rd Edition*, Morgan Kaufmann, 2011.
8. I. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques ,4th Edition*, Morgan Kaufmann, 2016.
9. Cathy O'Neil, and Rachel Schutt, *Doing Data Science, Straight Talk from The Frontline,* O'Reilly, 2014.

Important Notes:

(1) Students are expected to spend a total of 9 hours (i.e. 3 hours of class contact and 6 hours of personal study) per week to achieve the course learning outcomes.

(2) Students shall be aware of the University regulations about dishonest practice in course work, tests and examinations, and the possible consequences as stipulated in the Regulations Governing University Examinations. In particular, plagiarism, being a kind of dishonest practice, is "the presentation of another person's work without proper acknowledgement of the source, including exact phrases, or summarised ideas, or even footnotes/citations, whether protected by copyright or not, as the student's own work". Students are required to strictly follow university regulations governing academic integrity and honesty.

(3) Students are required to submit writing assignment(s) using Turnitin.

(4) To enhance students' understanding of plagiarism, a mini-course "Online Tutorial on Plagiarism Awareness" is available on https://pla.ln.edu.hk/.

Rubric for Midterm test of CDS4001 - Best Practices of Data Science

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Recognize the fundamental principles and practical applications of data science in different domains	The student demonstrates a clear understanding of the practical applications in different domains. The student can elaborate nearly all practical applications. The elaborations are completely / nearly completely correct and precise.	The student demonstrates a reasonable understanding of the practical applications in different domains. The student can elaborate most practical applications. Some elaborations are not completely correct and precise.	The student demonstrates limited understanding of the practical applications in different domains. The student can only elaborate a few practical applications.
Describe and effectively apply the data science process and techniques to problem solving	Correctly describes and applies nearly all data science processes and techniques to problem solving.	Correctly describes and applies most data science processes and techniques to problem solving. But some descriptions and applications are unclear or only partially correct.	Describes and applies a few data science processes and techniques to problem solving. However most descriptions and applications are unclear or incorrect.
Apply critical thinking skills to analyze data science problems and provide relevant data for decision making	Applies many critical thinking skills to analyze problems and provide data support for decision making. Most results and methods are correct.	Applies some critical thinking skills to analyze problems and provide data support for decision making. Some results and methods are correct.	Applies only a few critical thinking skills to analyze problems and provide data support for decision making. Most results and methods are incorrect.
Formulate the problems creatively and solve them using different algorithms and methods	Formulates problems creatively and solves them using many different algorithms and methods. Most result and methods are correct.	Formulates problems creatively and solves them using some different algorithms and methods. Some results and methods are correct.	Formulates only a few problems creatively and solves them using some different algorithms and methods. However, most result and methods are incorrect.
Evaluate and validate the result obtained in a data science process	Identifies many methods to evaluate and validate the results obtained in a data science process.	Identifies some methods to evaluate and validate the results obtained in a data science process.	The student does now know how to evaluate and validate the results obtained in a data science process.
Interpret the results and deliver the findings	Clearly and correctly delivers most findings and communicates with different stakeholders with a variety of diverse backgrounds.	Clearly delivers some findings and communicates with different stakeholders with a variety of diverse backgrounds. However, some findings are unclear or only partially incorrect.	Delivers the findings and communicates with different stakeholders with a variety of diverse backgrounds. However most findings are unclear and incorrect.

Rubric for Case project of CDS4001 - Best Practices of Data Science

Assessment Process:
On completion of the Case Project, each group will be evaluated using the following rubics for their project.

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Describe and effectively apply the data science process and techniques to problem solving	Correctly describe and apply nearly all data science processes and techniques to problem solving.	Correctly describe and apply most data science processes and techniques to problem solving. However some descriptions and applications are unclear or partially incorrect.	Describe and apply a few data science processes and techniques to problem solving. However most descriptions and applications are unclear or incorrect.
Apply critical thinking skills to analyze data science problems and provide relevant data for decision making	Apply many critical thinking skills to analyze problems and provide data support for decision making. Most result and methods are correct.	Apply some critical thinking skills to analyze problems and provide data support for decision making. Some results and methods are correct.	Apply only a few critical thinking skills to analyze problems and provide data support for decision making. And most results and methods are incorrect.
Formulate the problems creatively and solve them using different algorithms and methods	Formulate problems creatively and solve them using many different algorithms and methods. Most results and methods are correct.	Formulate problems creatively and solve them using some different algorithms and methods. Many results and methods are correct.	Formulate only a few problems creatively and solve them using some different algorithms and methods. However, most results and methods are incorrect.
Evaluate and validate the result obtained in a data science process	Identify many methods to evaluate and validate the result obtained in a data science process.	Identify some methods and manner to evaluate and validate the result obtained in a data science process.	The students do now know how to evaluate and validate the results obtained in a data science process.
Interpret the results and deliver the findings	Clearly and correctly deliver most findings and communicate with different stakeholders with a variety of diverse backgrounds.	Clearly deliver many findings and communicate with different stakeholders with a variety of diverse backgrounds. However, some findings are unclear or partially correct.	Deliver the findings and communicate with different stakeholders with a variety of diverse backgrounds. However, most findsing are unclear and incorrect.
Implement programs based on various data structures and object-oriented programming	Students can correctly select the right data structure(s) to handle the given problem(s). The implementations are correct, simple, and straight forward.	Students can correctly select the right data structure(s) to handle the given problem(s). The implementations are nearly correct. Unnecessary codes are used in the implementations.	Students select the wrong data structure(s) to handle the given problem(s), or the implementations are basically incorrect.
Develop and deploy the database management system to solve some real-world problem	Correctly develop and deploy the database management system to solve some real-world problems. Most results and methods are correct.	Correctly develop and deploy the database management system to solve some real-world problems. However, some results and methods are incorrect.	Incorrectly develop and deploy the database management system to solve some real-world problems, andmost results and methods are incorrect.

Rubric for Assignments of CDS4001 - Best Practices of Data Science

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Describe and effectively apply the data science process and techniques to problem solving	Correctly describes and applies nearly all data science process and techniques to problem solving	Correctly describes and applies most data science process and techniques to problem solving. But some descriptions and applications are unclear or only partially correct.	Describes and applies a few data science process and techniques to problem solving. However, most descriptions and applications are unclear or incorrect.
Apply critical thinking skills to analyze data science problems and provide relevant data for decision making	Applies many critical thinking skills to analyze problems and provide data support for decision making. Most result and methods are correct.	Applies some critical thinking skills to analyze problems and provide data support for decision making. Some results and methods are correct.	Applies only a few critical thinking skills to analyze problems and provide data support for decision making. And most results and methods are incorrect.
Formulate the problems creatively and solve them using different algorithms and methods	Formulates problems creatively and solve them using many different algorithms and methods. Most result and methods are correct.	Formulates problems creatively and solve them using some different algorithms and methods. Many results and methods are correct.	Formulates only a few problems creatively and solve them using some different algorithms and methods However, most result and methods are incorrect.
Evaluate and validate the result obtained in a data science process	Identifies many methods to evaluate and validate the results obtained in a data science process.	Identifies some methods to evaluate and validate the results obtained in a data science process.	The student does now know how to evaluate and validate the results obtained in a data science process.
Develop and deploy the database management system to solve some real-world problem	Correctly develops and deploys the database management system to solve some real-world problems. Most results and methods are correct.	Correctly develops and deploys the database management system to solve some real-world problems. However, some results and methods are incorrect.	Incorrectly develops and deploys the database management system to solve some real-world problems. And most results and methods are incorrect.

Rubric for Class Attendance and Participation of CDS4001 - Best Practices of Data Science

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Recognize the fundamental principles and practical applications of data science in different domains	The student demonstrates a clear understanding of the practical applications in different domains. The student can elaborate nearly all practical applications. The elaborations are completely / nearly completely correct and precise.	The student demonstrates a reasonable understanding of the practical applications in different domains. The student can elaborate most practical applications. Some elaborations are not completely correct and precise.	The student demonstrates limited understanding of the practical applications in different domains. The student can only elaborate a few practical applications.
Describe and effectively apply the data science process and techniques to problem solving	Correctly describes and applies nearly all data science process and techniques to problem solving	Correctly describes and applies most data science process and techniques to problem solving. But some descriptions and applications are unclear or only partially correct.	Describes and applies a few data science process and techniques to problem solving. But most descriptions and applications are unclear or incorrect.
Interpret the results and deliver the findings	Clearly and correctly delivers most findings and communicates with different stakeholders with a variety of diverse backgrounds.	Clearly delivers some findings and communicates with different stakeholders with a variety of diverse backgrounds. However, some findings are unclear or only partially correct.	Delivers the findings and communicates with different stakeholders with a variety of diverse backgrounds. However, most findings are unclear and incorrect.