Lingnan University

Department of Computing and Decision Sciences

Course Syllabus

Course Title	:	Data Mining
Course Code	:	CDS3004
Recommended Study Year	:	3
No. of Credits/Term	:	3
Mode of Tuition	:	Sectional Approach
Class Contact Hours	:	3 hours per week
Category in Major Prog.	:	Required
Prerequisite(s)	:	CDS2002 Introduction to Artificial Intelligence
Co-requisite	:	Nil
Exclusion	:	Nil
Exemption Requirement	:	Nil

Brief Course Description:

Data mining is an important component of data science that discovers knowledge from huge databases. Data mining is an interdisciplinary field, integrating statistics, pattern recognition, neuro-computing, machine learning and databases. It is also one of the fundamentals to extracting interesting knowledge (domain-specific rules, patterns, constraints and regularity). Students will learn the basic principles and core ideas of data mining. This course also covers many data mining approaches to discovering knowledge from a vast amount of valuable databases in business, finance, urban and medicine. Quantitative analytical skills are taught to interpret data mining models. Current IT skills are also covered in the course.

Aims:

This course aims at describing the whole data science process, the concepts of data mining, and the practical applications in the domains of Science, Social Science, Arts, and Business. It covers useful tools for the analysis, understanding and extraction of useful information and knowledge from huge real world databases in business, finance, urban, and medicine. Students will study and participate in the workflow of the data science process. Topics will include problem understanding, data understanding, data curation, data preprocessing, clustering, classification, model evaluation, visualization tools, and model deployment. In particular, cases studies will be given to show how to extract and present implicit, previously unknown, and potentially useful information from the results of the statistical and machine learning algorithms, such as k-mean clustering, random forest, Bayesian networks, and neural networks. Software packages such as WEKA, IBM SPSS Modeler, SAS Enterprise Miner, Microsoft Azure ML and/or PandasAI will be the major tools for the whole data mining process.

Learning Outcomes (LOs):

Upon the successful completion of this course, the student will be able to:

1. Identify the practical applications of data mining in different domains; (PLO1)
2. Identify the whole data science process; (PLO2)
3. Recognize different data mining techniques for various data; (PLO5)

4. Apply different data mining techniques for various problems; (PLO5)
5. Collect, store, process, and visualize data properly; (PLO6)
6. Evaluate and validate the results produced by data mining algorithms; (PLO7)
7. Interpret the results and deliver the findings; (PLO8)

Indicative Contents:

Data Mining
Basic concepts and the practical applications
The whole data science process: problem understanding, data understanding, data preprocessing, modelling, evaluation and validation, deployment
Different types of knowledge

Data Curation, Data Preprocessing, and Data Types
Questionnaire, web crawler, data imputation, data cleansing, data integration, data transformation, numerical data type, categorical data type, ordinal data type, feature selection, automatic feature engineering using Artificial Intelligence (AI) and generative AI, intelligent data visualization

Classification
Random tree, random forest, rule-based classifiers, nearest-neighbor classifiers, Bayesian classifiers, association rule mining, ensemble methods

Clustering
k-mean clustering, hierarchical clustering, density-based spatial clustering of applications with noise

Evaluation and Validation
Introduction to evaluation metrics, cross validation, nested cross validation, model evaluation using generative AI

Reporting
Sensitivity analysis, visualization, and precision-recall curve

Teaching Method:

The concepts and principles of data mining and its applications in different fields will be covered in lectures. The whole data science process and the tasks performed in different steps will also be discussed in lectures. Data mining software packages such as WEKA, IBM SPSS Modeler, SAS Enterprise Miner, Microsoft Azure ML, and/or PandasAI will be taught during the laboratories. A case study will be carried out throughout the course to demonstrate the process of applying different procedures and data mining techniques to discover knowledge from a huge database with more than 100,000 records and many variables. Students are required to perform a group project to apply the concepts and principles covered in this course to induce useful and valid knowledge and/or patterns from other database with many cases and variables. They are required to present their findings and the business implications derived from the project results.

Measurement of Learning Outcomes:

	Class Attendance and Participation	Assignments	Group Project	Examination
1. Identify the practical applications of data mining in different domains	x			x
2. Identify the whole data science process	x			x
3. Recognize different data mining techniques for various data		x	x	x
4. Apply different data mining techniques for various problems		x	x	
5. Collect, store, process, and visualize data properly		x	x	
6. Evaluate and validate the results produced by data mining algorithms		x	x	x
7. Interpret the results and deliver the findings			x	

1. There are a number of classroom activities to evaluate if the students can identify different practical applications and the whole data science process (LO1, LO2, PLO1, and PLO2).
2. Assignments require students to demonstrate their understandings of different data mining techniques, apply different data mining techniques, collect data, store data, process data, visualize data, evaluate and validate the results (LO3-L06, PLO5-PLO7).
3. The group project requires students to demonstrate their understandings of different data mining techniques and apply different data mining techniques provided in different software packages. Students are required to apply IT skills to collect, store, process, and visualize data. The appropriate procedures and methods should be used to evaluate and validate the results produced by different data mining algorithms. They are also required to interpret the results and deliver the findings via project oral presentation (LO3-7, PLO5-8).
4. The examination can evaluate if the students can understand different practical applications and describe the whole data science process. Students are required to demonstrate their understandings of different data mining techniques, evaluation methods, and validation procedure (LO1-3, LO6, PLO1-2, PLO5, and PLO7).

Assessment:

Class Attendance and Participation	5%
Assignments	30%
Group Project	30%
Final Examination	35%
Total	100%

Required/Essential Readings:

1. Tan, Pang-Ning, Steinbach, Michael, and Kumar, Vipin. *Introduction to Data Mining*, Pearson Education Dorling Kindersley, 2016.
2. Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition*, Morgan Kauffmann, 2016.

Recommended/Supplementary Readings:

1. Baesens, B., *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*, Wiley, 2014.
2. Bartlett, R., *A Practitioner's Guide to Business Analytics: Using Data Analysis Tools to Improve Your Organization's Decision Making and Strategy*, McGraw-Hill, 2013.
3. Foreman, J. W., *Data Smart: Using Data Science to Transform Information into Insight*, Wiley, 2013.
4. Han, J., Kamber, M., and Pei, J., *Data Mining: Concepts and Techniques*, *3rd Edition*, Morgan Kauffmann, 2011.
5. Hansen, D., Shneiderman, B., and Smith, M. A., *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*, Morgan Kauffmann, 2010.
6. Howson, C., *Successful Business Intelligence: Unlock the Value of BI & Big Data*, McGram-Hill, 2013.
7. Linoff, G. S. and Berry, M., *Data Mining Techniques: For Marketing, Sales, and Customer Relation- ship Management*, *3rd Edition*, Wiley, 2011.
8. Maisel, L. and Cokins, G., *Predictive Business Analytics: Forward Looking Capabilities to Improve Business Performance*, Wiley, 2014.
9. Provost, F. and Fawcett, T., *Data Science for Business: What you need to know about data mining and data-analytic thinking,* O'Reilly Media, 2013.
10. Siegel, E. and Davenport, T. H., *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*, *2nd Edition*, Wiley, 2016.

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 Final Examination of CDS3004 - Data Mining

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Recognize the practical applications in different domains (LO1)	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 the whole data science process (LO2)	The student can describe all procedures in the whole process correctly.	The student can describe most procedures in the whole process correctly.	The student can only describe a few procedures in the whole process correctly.
Recognize different data mining techniques for various data (LO3)	The student gives correct descriptions of most data mining techniques.	The student gives correct descriptions of some data mining techniques.	The student gives correct descriptions of few data mining techniques.
Evaluate and validate the results produced by data mining algorithms (LO6)	**All** of the following points are achieved ▪ Employ the right validation procedure(s) ▪ Use the right evaluation metrics ▪ Calculate all measurement values correctly.	**Two** of the following points are achieved ▪ Employ the right validation procedure(s) ▪ Use the right evaluation metrics ▪ Calculate all measurement values correctly.	**None or One** of the following tasks are achieved ▪ Employ the right validation procedure(s) ▪ Use the right evaluation metrics ▪ Calculate all measurement values correctly.
Presentation	Content of submission/ presentation is well formatted with a clearly readable layout and no/very few grammatical/ mistakes.	Content of submission/ presentation is properly formatted with a reasonable layout and no more than a few grammatical mistakes.	Content of submission/ presentation is not properly formatted and/or there are more than a few grammatical mistakes.

Rubric for Individual Assignments of CDS3004 - Data Mining

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Able to load the datasets, assign feature characteristics, handle missing and noisy data (LO5)	**All** of the following points are achieved ▪ Load large datasets ▪ Assign the right characteristics to all features ▪ Handle all missing and noisy data correctly	**Two** of the following points are achieved ▪ Load large datasets ▪ Assign the right characteristics to all features ▪ Handle all missing and noisy data correctly	**None or One** of the following points are achieved ▪ Load large datasets ▪ Assign the right characteristics to all features ▪ Handle all missing and noisy data correctly
Able to apply Data Mining techniques to select the relevant features (LO5)	Can use the appropriate feature selection methods to select most of the relevant features and remove most of the irrelevant features.	More than 5% of the relevant features have not been selected **or** more than 5% of the irrelevant features have not been removed.	More than 10% of the relevant features have not been selected **or** more than 10% of the irrelevant features have not been removed.
Can employ Data Mining techniques to perform advanced feature transformations (LO5)	Appropriate feature transformation methods have been performed. Suitable new features have been generated. The settings of the methods are completely correct.	Appropriate feature transformation methods have been performed, but the transformed features have some minor issues, because the settings of the methods are not completely correct.	Feature transformations have not been performed or most feature transformations have been done incorrectly.
Can perform data visualization (LO5)	Appropriate data visualization methods have been performed. Suitable visualization results have been obtained. The settings of the methods are completely correct.	Appropriate data visualization methods have been performed, but the results have some minor issues, because the settings of the methods are not completely correct.	Data visualization has not been performed or most data visualizations have been done incorrectly.
Recognize different Data Mining techniques for various datasets (LO3)	Appropriate data mining methods have been used for all datasets.	Appropriate data mining methods have been partially used for most datasets.	Inappropriate data mining methods have been used for most datasets.
Able to use Data Mining techniques to learn models from data (LO4)	The procedure of using the data mining methods and the settings of these methods are completely correct. The learnt models are accurate.	There are a few errors in the procedure of using the data mining methods and the settings of these methods. Thus the learnt models contain some flaws.	The procedure of using the data mining methods and the settings of these methods cannot be developed or contains major errors.Thus, the learnt models are incorrect.
Able to use the right techniques to evaluate the performance of the learnt models (LO6)	Appropriate methods (e.g. cross-validation, nested cross-validation) have been applied to divide the dataset into training, testing and/or validation data sets. The models using the training and/or the validation datasets have been learnt and validated. Suitable methods have been applied to evaluate the performance of the model on the testing dataset.	Performance evaluation has been done. The evaluation strategy is generally correct. But the evaluation results are biased or unstable.	Performance evaluation has not been done or the evaluation strategy is incorrect.

Rubric for Group Project Presentation of CDS3004 - Data Mining

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Appropriate time allocation and pace.	Allocate time appropriately, and managed time effectively, with smooth progression. Appropriate pace. Start presentation punctually.	Marginally long or marginally short but uses time reasonably effectively. Reasonable pace. Start presentation relatively punctually.	Significantly too short or too long and did not use time effectively. Pace is significantly too fast or too slow. Don't start presentation punctually.
Clear, logically organized and relevant content.	Information included is always relevant. Clearly stated and developed points. Material flows extremely well and is well organized. No ambiguities are left unexplained.	Information included is generally relevant. Key points are relatively clear. Most information presented in logical sequence; sufficiently well-organized with generally satisfactory flow. Some ambiguities are left unexplained.	Much of the information included is not relevant and even key points are not clear. Presentation is choppy or disjointed, does not flow well, and has no apparent logical order.
Effective use of presentation tools.	Balanced and proper use of presentation tools with little or no distraction (e.g., appropriate animation/pictures, appropriate information on one slide, good color combination, clear titles, etc.)	Generally good use of presentation tools. Some distractions but they are not overwhelming (e.g., reasonable animation/pictures, fair information on one slide, fair color combination, fair titles, etc.)	Poor use of presentation tools and/or many distractions (e.g., too much animation/pictures, too much information on one slide, poor color combination, absence of titles, etc.)
Uses good body language, eye contact, appropriate voice tone.	Shows poise and composure; makes good eye contact with audience; balanced posture; shows enthusiasm and confidence; uses voice tone effectively.	Fairly poised and composed; makes fairly good eye contact with audience; balanced posture; shows some enthusiasm and confidence; uses voice tone relatively effectively.	Little poise and composure; makes little or no eye contact with audience; poor posture; shows little or no enthusiasm and confidence; uses voice tone ineffectively or too monotonously.
Gains/holds attention	Provides good motivation to engage the audience's interest. Presents the content in a manner that captivates the audience's attention.	Provides reasonable motivation to engage the audience's interest. Audience is reasonably engaged but there are instances where the presentation is otherwise dull.	Provides insufficient motivation to engage the audience's interest. Dull presentation of content that does not engage the audience.
Uses instructor defined role appropriate dress	Professionally dressed as expected by the instructor.	Minor deviations from instructor's expectations.	Do not dress in a manner expected by the instructor.
Clarity of speech/Accuracy of grammar & pronunciation	Voice is consistently comprehensible; grammar and pronunciation are accurate.	Voice is generally comprehensible; grammar and pronunciation are adequate but with some mistakes.	Voice is incomprehensible on several occasions; many mistakes in terms of grammar and pronunciation.

Rubric for Group Project of CDS3004 - Data Mining

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Problem definition (demonstrate the understanding of the problem and consider alternative solutions)	Clearly state the problem, list related constrains, and identify alternative solutions.	The problem is stated but related constrains and alternative solutions are not considered thoroughly.	The problem is marginally defined and with no consideration of constrains and alternative solutions.
Solution design	The design of the solution is strongly related to the problem and clearly explained the approaches and techniques involved. Moreover, the design of the solution is very innovative.	The design of the solution is related to the problem but not clearly explained the approach and techniques involved. However, the design of the solution is not very innovative.	The design of the solution is weakly related to the problem and no explanation of the approach and technique involved.
Able to collect datasets, load the datasets, assign feature characteristics, handle missing and noisy data (LO5)	**All** of the following points are achieved ▪ Collect datasets ▪ Load large datasets ▪ Assign the right characteristics to all features ▪ Handle all missing and noisy data correctly	**Two** of the following points are achieved ▪ Collect datasets ▪ Load large datasets ▪ Assign the right characteristics to all features ▪ Handle all missing and noisy data correctly	**None or One** of the following points are achieved ▪ Collect datasets ▪ Load large datasets ▪ Assign the right characteristics to all features ▪ Handle all missing and noisy data correctly
Able to apply Data Mining techniques to select the features (LO5)	Can use the appropriate feature selection methods to select most of the relevant features and remove most of the irrelevant features.	A few relevant features have not been selected **or** a few irrelevant features have not been removed.	Several relevant features have not been selected **or** several irrelevant features have not been removed.
Can employ Data Mining techniques to perform advanced feature transformations (LO5)	Appropriate feature transformation methods have been performed. Suitable new features have been generated. The settings of the methods are completely correct.	Appropriate feature transformation methods have been performed, but the transformed features have some minor issues, because the settings of the methods are not completely correct.	Feature transformations have not been performed or most feature transformations have been done incorrectly.
Can perform data visualization (LO5)	Appropriate data visualization methods have been performed. Suitable visualization results have been obtained. The settings of the methods are completely correct.	Appropriate data visualization methods have been performed to a large extent, but the results have some minor issues, because the settings of the methods contains a few errors.	Data visualization is not been performed appropriately with many errors or not been performed. Most data visualizations have been done incorrectly.
Appy different Data Mining techniques for various data (LO3)	Appropriate data mining methods have been used consistently for all datasets.	Appropriate data mining methods have been used for most datasets most of the time with some minor errors.	Partially to no usage of appropriate data mining methods.
Able to use Data Mining techniques to learn models from data (LO4)	The procedure of using the data mining methods and the settings of these methods are completely correct. The learnt models are accurate.	There are a few to some minor errors in the procedure of using the data mining methods and the settings of these methods are not completely correct. Thus the learnt models have some issues.	There are several errors in the procedure of using the data mining methods and the settings of these methods are mostly faulty. The learnt models are of minmal satisfactory.

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Able to use the right techniques to evaluate the performance of the learnt models (LO6)	Appropriate methods (e.g. cross-validation, nested cross-validation) have been applied to divide the dataset into training, testing and/or validation data sets. Learn and validate the models using the training and/or the validation datasets. Apply the suitable methods to evaluate the performance of the model on the testing dataset.	Performance evaluation has been done. The evaluation strategy is generally correct with a few to some minor errors. The evaluation results are biased or unstable to a small extent.	Performance evaluation has been done with many errors or entirely wrong. .The evaluation strategy is incorrectly chosen and biased or unstable to a large extent.
Able to interpret the results (LO7)	The interpretation is completely correct and insightful. Critical evaluation of the limitations is correct.	There are some issues in the interpretation of the results. Critical evaluation of the limitations is not completely correct.	The Interpretation of the results is incorrect. Critical evaluation of the limitations is not provided.
Able to deliver the learnt models (LO7)	Apply the learnt models for decision making on new examples. The decisions are sound and reasonable.	Apply the learnt models for decision making on new examples to a large extent with a few to some minor errors. The decisions are not completely sound and reasonable.	Only a limited to a very confined extent of application of the learnt models for decision making to new examples. The decisions are of a large extent to completely unconvincing and unreasonable.

Rubric for Class Attendance and Participation of CDS3004 - Data Mining

Criteria	Very good (4-6)	Satisfactory (2-4)	Unsatisfactory (0-2)
Attendance	Full, punctual attendance in class and mandatory seminars	Occasional absences or lateness from class or mandatory seminars	Frequent or recurring absence or lateness from class or mandatory seminars
Class Participation	Active class participation and leadership in group activities.	Passive class participation and active in group activities.	Lack of participation and active disruption of class and group activities.
Recognize the practical applications in different domains (LO1)	The student can describe nearly all practical applications.	The student can describe some or most practical applications. Some descriptions are not completely correct and precise.	The student can only describe a few practical applications.
Describe the whole data science process (LO2)	The student can describe all procedures in the whole process correctly.	The student can describe some or most procedures in the whole process correctly.	The student can only describe a few procedures in the whole process correctly.