

Subject Description Form

Subject Code	DSAI5101
Subject Title	Statistical Data Mining
Credit Value	3
Level	5
Pre-requisite/ Co-requisite/ Exclusion	Pre-requisite: AMA528 or AMA563 or DSAI5102
Objectives	To provide a comprehensive introduction to data mining. To present fundamental concepts and algorithms for each topic, to provide the students with the necessary background for the application of data mining to real problems, and to provide a starting point for students who are interested in pursuing research in data mining or related fields.
Intended Learning Outcomes	Upon completion of the subject, students will be able to: (a) Understand the principle of data mining; (b) Apply data mining techniques; (c) Report and interpret results in scientific manner; (d) Apply statistical package in data mining.
Subject Synopsis/ Indicative Syllabus	<p><u>Preliminaries:</u> Vectors and matrices; eigenvalues and eigenvectors; singular value decomposition; expectation and variance; point estimation; data types, data quality.</p> <p><u>Regression and Classification:</u> Linear regression, least squares methods; normal distributions, error analysis for least squares, the Gauss-Markov theorem; model overfitting; subset selection, shrinkage methods, ridge regression, the LASSO; dimension reduction, principal component analysis; kernel methods, kernel ridge regression; nearest neighbor methods; expectation-maximization; logistic regression; decision trees; naive Bayes classifiers; random forests; support vector machines; artificial neural networks; evaluating the performance of classifiers; receiver operating characteristic curves, AUC scores</p> <p><u>Clustering and Association Analysis:</u> Clustering analysis; hierarchical clustering; prototype-based clustering, density-based clustering, graph-based clustering; k-means clustering; spectral clustering; Association analysis</p> <p><u>Data Mining in Practice:</u> Data processing and measurements; introduction to selected software packages for data mining; data visualization; implementation of selected algorithms</p>

Teaching/Learning Methodology	The subject will be delivered mainly through lectures and tutorials. The teaching and learning approach is mainly problem-solving oriented. The approach aims at the development of data mining techniques and how the techniques can be applied to problem solving. Students are encouraged to adopt a deep study approach by employing high level cognitive strategies, such as critical and evaluative thinking, relating, integrating and applying theories to practice.																																						
Assessment Methods in Alignment with Intended Learning Outcomes	<table border="1"> <thead> <tr> <th rowspan="2">Specific assessment methods/tasks</th><th rowspan="2">% weighting</th><th colspan="4">Intended subject learning outcomes to be assessed (Please tick as appropriate)</th></tr> <tr> <th>a</th><th>b</th><th>c</th><th>d</th></tr> </thead> <tbody> <tr> <td>1. Assignments/Projects</td><td>20%</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td></tr> <tr> <td>2. Midterm Test</td><td>20%</td><td>✓</td><td>✓</td><td>✓</td><td></td></tr> <tr> <td>3. Examination</td><td>60%</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td></tr> <tr> <td>Total</td><td>100%</td><td colspan="4"></td></tr> </tbody> </table> <p>Explanation of the appropriateness of the assessment methods in assessing the intended learning outcomes:</p> <p>Practical use of statistical data mining with application of computer software is emphasized that 40% continuous assessment is appropriate in assessing students' performance.</p> <p>Continuous Assessment consists of projects of real life problems, assignments and a midterm test.</p> <p>The mathematical and statistical foundation of data mining algorithms are assessed in both Continuous Assessment and Examination.</p>					Specific assessment methods/tasks	% weighting	Intended subject learning outcomes to be assessed (Please tick as appropriate)				a	b	c	d	1. Assignments/Projects	20%	✓	✓	✓	✓	2. Midterm Test	20%	✓	✓	✓		3. Examination	60%	✓	✓	✓	✓	Total	100%				
Specific assessment methods/tasks	% weighting	Intended subject learning outcomes to be assessed (Please tick as appropriate)																																					
		a	b	c	d																																		
1. Assignments/Projects	20%	✓	✓	✓	✓																																		
2. Midterm Test	20%	✓	✓	✓																																			
3. Examination	60%	✓	✓	✓	✓																																		
Total	100%																																						
Student Study Effort Required	Class contact:																																						
	▪ Lectures/Tutorials		39 Hrs.																																				
	Other student study effort:																																						
	▪ Assignments/Projects		58 Hrs.																																				
	▪ Self-study		40 Hrs.																																				
	Total student study effort		137 Hrs.																																				

**Reading List and
References**Textbook:

Tan, P.N., Steinbach,
M., and Kumar, V.

Introduction to Data Mining Pearson 2006

References:

Giudici, P.

Applied Data Mining:
Statistical Methods For
Business And Industry

Wiley 2005

Han, J., Kamber, M.,
and Pei, J.

Data Mining: Concepts
and Techniques, 3rd
Edition.

Morgan Kaufmann,
2011

James, G., Witten, D.,
Hastie, T., and
Tibshirani, R.

An Introduction to
Statistical Learning

Springer 2013

Matignon, R.

Data Mining Using SAS
Enterprise Miner

Wiley 2007

Hastie, T., Tibshirani,
R, and Friedman, J.

The Elements of Statistical
Learning

Springer 2009

Refaat, M.

Data Preparation for Data
Mining Using SAS

Morgan Kaufmann,
2006

Johnson, R.A., and
Wichern, D.W.

Applied Multivariate
Statistical Analysis
6th edition

Pearson Prentice
Hall 2007