Subject Description Form

Subject Code	COMP4433
Subject Title	Data Mining and Data Warehousing
Credit Value	3
Level	3
Pre-requisite / Co-requisite / Exclusion	**Pre-requisite**: COMP2411 or equivalent introductory database subject
Objectives	This subject aims at equipping students with the latest knowledge and skills to: 1. create a clean, consistent repository of data within a data warehouse for large corporations; 2. utilise various techniques developed for data mining to discover interesting patterns in large databases; 3. use existing commercial or public-domain tools to perform data mining tasks to solve real problems in business and commerce; and 4. expose students to new techniques and ideas that can be used to improve the effectiveness of current data mining tools.
Intended Learning Outcomes	Upon completion of the subject, students will be able to: <u>*Professional/academic knowledge and skills*</u> (a) identify and analyse why there is a need for data warehouse in addition to traditional operational database systems, motivated by real examples; (b) conduct in-depth analysis of the key components in typical and advanced data warehouse architectures; (c) design a data warehouse and understand the process required to construct one; (d) identify and analyse why there is a need for data mining and in what ways it is different from traditional statistical techniques, motivated by real examples; (e) learn and master the algorithms made available by popular commercial data mining software; (f) solve real data mining problems by using the right tools to find interesting patterns; (g) obtain deep understanding of a typical knowledge discovery process; (h) obtain hands-on experience with some popular data mining software; <u>*Attributes for all-roundedness*</u> (i) apply data mining and data warehousing tools;

	(j) learn independently and search for relevant information to write reports to recommend appropriate data warehousing and data mining tools; and (k) generate innovative solutions individually or in groups and develop group work skills directly and indirectly.
Subject Synopsis/ Indicative Syllabus	**Topic** **1. Introduction to Data Warehousing and Data Mining** Introduction to data warehousing and data mining; possible application areas in business and finance; definitions and terminologies; types of data mining problems. **2. Data Warehousing** Data warehouse and data warehousing; data warehouse and the industry; definitions; operational databases vs. data warehouses. **3. Data Warehouse Architecture and Design** Data warehouse architecture and design; two-tier and three-tier architecture; star schema and snowflake schema; data characteristics; static and dynamic data; meta-data; data marts. **4. Data Replication and Online Analytical Processing** Data replication, data capturing and indexing, data transformation and cleansing; replicated data and derived data; Online Analytical Processing (OLAP); multidimensional databases; data cube. **5. Data Mining and Knowledge Discovery** Data mining and knowledge discovery, the data mining lifecycle; pre-processing; data transformation; types of problems and applications. **6. Association Rules** Mining of association rules; the Apriori algorithm; binary, quantitative and generalised association rules; interestingness measures. **7. Classification** Classification; decision tree based algorithms; Bayesian approach; statistical approaches, nearest neighbour approach; neural network based approach; genetic algorithms based technique; evaluation of classification model. **8. Clustering** Clustering; k-means algorithm; hierarchical algorithm; condorset; neural network and genetic algorithms based approach; evaluation of effectiveness. **9. Sequential Data Mining** Sequential data mining; time dependent data and temporal data; time series analysis; sub-sequence matching; classification and clustering of temporal data; prediction.

	10. Other Techniques Computation intelligence techniques; fuzzy logic, genetic algorithms and neural networks for data mining. Laboratory Experiment: **Topic** 1. Discover Association rules and sequential patterns using data mining tools 2. Discover Classification rules using data mining tools 3. Discover Clusters using data mining tools Case Study: 1. Application of data mining techniques to solve real business problems. 2. Attributes leading to success and failure of data warehousing projects tutorials when appropriate.
Teaching/ Learning Methodology	This subject consists mainly of class lectures and laboratory sessions. For the class lectures, various cases will be presented to help student understand why there is a need for data warehouse to be built and why data mining is important for modern day business intelligence. Students will be given time to participate in discussions when the cases are presented. All assignments and projects will also be given in the form of different cases collected so as to allow students to learn more about how data warehouse and data mining can be and have been used in real business environment. For the projects and assignments, students are expected to learn independently and think critically with minimise guidance. They are expected to practice their writing kills through project documentations and report writing. As students will work in teams on the project, they are expected to also learn to work with each other collaboratively. During laboratory sessions, students will be introduced to popular software products that can support the building of data warehouses and the mining of them. Students are expected to solve real data mining problems by using the right tools to find interesting patterns.

Assessment Methods in Alignment with Intended Learning Outcomes

Specific assessment methods/tasks	% weighting	Intended subject learning outcomes to be assessed										
		a	b	c	d	e	f	g	h	i	j	k
Continuous Assessment	**55%**											
1. Assignment		✓		✓	✓					✓	✓	
2. Project					✓	✓	✓	✓	✓	✓	✓	✓
Examination	**45%**	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Total	100 %											

The assessment consists of written assignments, a group project and an examination. For the assignments and projects, they are designed to ensure that students are able to achieve the learning outcomes intended for this subject. They are expected to tackle a number of cases drawn from different application areas in business and commerce so that they can understand why there is a need for data warehouse in addition to traditional operational database systems and why data mining is important for modern-day business intelligence. In addition, students will learn through the questions and cases, when a particular data warehouse architecture or when a particular data mining algorithm is useful and should be used. Questions in the assignments are expected to help students learning the details of the data mining algorithm and the use of popular data mining software. They are also expected to use such popular tool as Oracle Warehouse Builder to construct data warehouses. For the projects, students are expected to work in groups of three to four to tackle a real case involving the design of a data warehouse or the use of data mining to mine very large data bases. They are expected to learn how real-world problems in business and commerce should be tackled using real-world tools as Oracle's Warehouse Builder or IBM's Clementine data mining system. They are expected to learn independently and search for relevant information to write reports to recommend appropriate data warehousing and data mining tools. Students are expected to practice their writing skills with project document and report writing. They will learn to develop critical thinking and team work skills.

Student Study Effort Expected

Class contact:	
▪ Lectures/Laboratory	39 Hrs.
▪ Tutorials	0 Hrs.
Other student study effort:	
▪ Assignments and Case Studies	45 Hrs.
▪ Projects and Research	25 Hrs.
Total student study effort	109 Hrs.

Reading List and References	**Reference Books:** 1. Han, Jiawei and Kamber, Micheline, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann, 2012. 2. Golfarelli, Matteo and Rizzi, Stefano, *Data Warehouse Design: Modern Principles and Methodologies*, McGraw-Hill, 2009. 3. Inmon, W.H., Strauss, Derek and Neushloss, Genia, *DW 2.0: The Architecture for the Next Generation of Data Warehousing*, Morgan Kaufmann, 2008. 4. Rokach, Lior and Maimon, Oded Z., *Data Mining with Decision Trees: Theory and Applications*, World Scientific, 2008. 5. Witten, Ian H., Frank, Eibe and Hall, Mark A., *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd Edition, Morgan Kaufmann, 2011. 6. Westphal, Christopher, *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*, CRC Press, 2008. 7. Cox, Earl, *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*, Morgan Kaufmann, 2005. 8. Liu, Bing, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, Berlin Heidelberg, 2009. 9. Tsiptsis, Konstantinos K. and Chorianopoulos, Antonios, *Data Mining Techniques in CRM: Inside Customer Segmentation*, Wiley, 2010. 10. Shapiro, A.F. and Jain, L.C., *Intelligent and Other Computational Techniques in Insurance: Theory and Applications*, World Scientific, 2003.