INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal, Hyderabad -500 043

INFORMATION TECHNOLOGY

COURSE DESCRIPTOR

Course Title	DATA WAREHOUSING AND DATA MINING				
Course Code	AIT006				
Programme	B. Tech				
Semester	VI	CSE	IT		
Course Type	Core				
Regulation	IARE - R16				
Course Structure	**Theory**			**Practical**	
	Lectures	**Tutorials**	**Credits**	**Laboratory**	**Credits**
	3	1	4	3	2
Chief Coordinator	Mr. Ch Suresh Kumar Raju, Assistant Professor				
Course Faculty	Mr. A Praveen, Assistant Professor Mr. Ch Suresh Kumar Raju, Assistant Professor				

I. COURSE OVERVIEW:

The course addresses the concepts, skills, methodologies, and models of data warehousing. The proper techniques for designing data warehouses for various business domains, and covers concepts for potential uses of the data warehouse and other data repositories in mining opportunities are addressed. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge- driven decisions.

II. COURSE PRE-REQUISITES:

Level	Course Code	Semester	Prerequisites	Credits
UG	ACS005	III	Database Management Systems	4
UG	AHS010	II	Probability and statistics	4

III. MARKS DISTRIBUTION:

Subject	SEE Examination	CIA Examination	Total Marks
Data Warehousing and Data Mining	70 Marks	30 Marks	100

IV. DELIVERY / INSTRUCTIONAL METHODOLOGIES:

✔	Chalk & Talk	✔	Quiz	✔	Assignments	✘	MOOCs
✔	LCD / PPT	✔	Seminars	✘	Mini Project	✘	Videos
✘	Open Ended Experiments						

V. EVALUATION METHODOLOGY:

The course will be evaluated for a total of 100 marks, with 30 marks for Continuous Internal Assessment (CIA) and 70 marks for Semester End Examination (SEE). Out of 30 marks allotted for CIA during the semester, marks are awarded by taking average of two CIA examinations or the marks scored in the make-up examination.

Semester End Examination (SEE): The SEE is conducted for 70 marks of 3 hours duration. The syllabus for the theory courses is divided into five units and each unit carries equal weightage in terms of marks distribution. The question paper pattern is as follows. Two full questions with "either" or "choice" will be drawn from each unit. Each question carries 14 marks. There could be a maximum of two sub divisions in a question.

The emphasis on the questions is broadly based on the following criteria:

50 %	To test the objectiveness of the concept.
50 %	To test the analytical skill of the concept OR to test the application skill of the concept.

Continuous Internal Assessment (CIA):
CIA is conducted for a total of 30 marks (Table 1), with 25 marks for Continuous Internal Examination (CIE), 05 marks for Quiz/ Alternative Assessment Tool (AAT).

Table 1: Assessment pattern for CIA

Component	**Theory**		**Total Marks**
Type of Assessment	CIE Exam	Quiz / AAT	
CIA Marks	25	05	30

Continuous Internal Examination (CIE):
Two CIE exams shall be conducted at the end of the 8th and 16th week of the semester respectively. The CIE exam is conducted for 25 marks of 2 hours duration consisting of two parts. Part–A shall have five compulsory questions of one mark each. In part–B, four out of five questions have to be answered where, each question carries 5 marks. Marks are awarded by taking average of marks scored in two CIE exams.

Quiz / Alternative Assessment Tool (AAT):
Two Quiz exams shall be online examination consisting of 25 multiple choice questions and are be answered by choosing the correct answer from a given set of choices (commonly four). Marks shall be awarded considering the average of two quizzes for every course. The AAT may include seminars, assignments, term paper, open ended experiments, five minutes video and MOOCs.

VI. HOW PROGRAM OUTCOMES ARE ASSESSED:

Program Outcomes (POs)		Strength	Proficiency assessed by
PO 1	**Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.	3	Presentation on real-world problems
PO 2	**Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences	3	Assignment
PO 3	**Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.	2	Assignment
PO 4	**Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.	2	Seminar
PO 5	**Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.	1	Seminar

3 = High; 2 = Medium; 1 = Low

VII. HOW PROGRAM SPECIFIC OUTCOMES ARE ASSESSED:

Program Specific Outcomes (PSOs)		Strength	Proficiency assessed by
PSO 1	**Professional Skills:** The ability to understand, analyze and develop computer programs in the areas related to algorithms, system software, multimedia, web design, big data analytics, and networking for efficient design of computer-based systems of varying complexity.	2	Seminar
PSO 2	**Software Engineering Practices**: The ability to apply standard practices and strategies in software service management using open-ended programming environments with agility to deliver a quality service for business success.	-	-
PSO 3	**Successful Career and Entrepreneurship:** The ability to employ modern computer languages, environments, and platforms in creating innovative career paths to be an entrepreneur, and a zest for higher studies.	1	Seminar

3 = High; 2 = Medium; 1 = Low

VIII. COURSE OBJECTIVES :

The course should enable the students to:	
I	Identifying necessity of Data Mining and Data Warehousing for the society.
II	Familiar with the process of data analysis, identifying the problems, and choosing the relevant models and algorithms to apply.
III	Develop skill in selecting the appropriate data mining algorithm for solving practical problems.
IV	Develop ability to design various algorithms based on data mining tools.
V	Create further interest in research and design of new Data Mining techniques and concepts.

IX. COURSE LEARNING OUTCOMES (CLOs):

COs	Course Outcome	CLOs	Course Learning Outcome
CO 1	Understand the concept of data warehouse and its architecture, components.	CLO 1	Learn data warehouse principles and find the differences between relational databases and data warehouse
		CLO 2	Explore on data warehouse architecture and its components
		CLO 3	Learn Data warehouse schemas
		CLO 4	Differentiate different OLAP Architectures
CO 2	Understand Data Mining concepts and knowledge discovery process	CL05	Understand Data Mining concepts and knowledge discovery process
		CLO 6	Explore on Data preprocessing techniques
		CLO 7	Apply task related attribute selection and transformation techniques
		CLO 8	Understand the Association rule mining problem
CO 3	Explore on decision tree construction and attribute selection	CLO 9	Illustrate the concept of Apriori algorithm for finding frequent items and generating association rules. Association rules.
		CL10	Illustrate the concept of Fp-growth algorithm and different representations of frequent item sets.
		CLO 11	Understand the classification problem and prediction
		CLO 12	Explore on decision tree construction and attribute selection
CO 4	Understand the classification problem and Bayesian classification	CLO 13	Understand the classification problem and Bayesian classification
		CLO 14	Illustrate the rule based and back propagation classification algorithms
		CLO 15	Understand the Cluster and Analysis.
		CLO 16	Understand the Types of data and categorization of major clustering methods.
CO 5	Explore on different hierarchical based methods, grid based and Model based methods.	CLO 17	Explore on partition algorithms for clustering.
		CLO 18	Explore on different hierarchical based methods, different density based methods, grid based and Model based methods.
		CLO 19	Understand the outlier Analysis.
		CLO 20	Understand mining complex data types.

X. COURSE LEARNING OUTCOMES (CLOs):

CLO Code	CLO's	At the end of the course, the student will have the ability to:	PO's Mapped	Strength of Mapping
AIT006.01	CLO 1	Learn data warehouse principles and find the differences between relational databases and data warehouse	PO 1	3

CLO Code	CLO's	At the end of the course, the student will have the ability to:	PO's Mapped	Strength of Mapping
AIT006.02	CLO 2	Explore on data warehouse architecture and its components	PO 1;PO4	2
AIT006.03	CLO 3	Learn Data warehouse schemas	PO 1;PO 2	2
AIT006.04	CLO 4	Differentiate different OLAP Architectures	PO1;PO2	2
AIT006.05	CLO 5	Understand Data Mining concepts and knowledge discovery process	PO 2;PO3,	2
AIT006.06	CLO 6	Explore on Data preprocessing techniques	PO 1; PO5	2
AIT006.07	CLO 7	Apply task related attribute selection and transformation techniques	PO 2;PO5	1
AIT006.08	CLO 8	Understand the Association rule mining problem	PO 2	2
AIT006.09	CLO 9	Illustrate the concept of Apriori algorithm for finding frequent items and generating association rules. Association rules.	PO 1;PO 3	2
AIT006.10	CLO 10	Illustrate the concept of Fp-growth algorithm and different representations of frequent item sets.	PO 1	3
AIT006.11	CLO 11	Understand the classification problem and prediction	PO 3;PO5	1
AIT006.12	CLO 12	Explore on decision tree construction and attribute selection	PO3	2
AIT006.13	CLO 13	Understand the classification problem and Bayesian classification	PO 3;PO 4	2
AIT006.14	CLO 14	Illustrate the rule based and back propagation classification algorithms	PO 3;PO 4	2
AIT006.15	CLO 15	Understand the Cluster and Analysis.	PO 3	2
AIT006.16	CLO 16	Understand the Types of data and categorization of major clustering methods.	PO 2;PO 3	2
AIT006.17	CLO 17	Explore on partition algorithms for clustering.	PO 2;PO 3	2
AIT006.18	CLO 18	Explore on different hierarchical based methods, different density based methods, grid based and Model based methods.	PO2;PO3	2
AIT006.19	CLO 19	Understand the outlier Analysis.	PO 3;PO5	1
AIT006.20	CLO 20	Understand mining complex data types.	PO 1;PO 2	2

3 = High; 2 = Medium; 1 = Low

XI. MAPPING COURSE OUTCOMES LEADING TO THE ACHIEVEMENT OF PROGRAM OUTCOMES AND PROGRAM SPECIFIC OUTCOMES

Course Outcomes (COs)	Program Outcomes and Program Specific Outcomes						
	PO 1	PO 2	PO 3	PO 4	PO 5	PSO 1	PSO 3
CO 1	3	2		2		3	1

Course Outcomes (COs)	Program Outcomes and Program Specific Outcomes						
	PO 1	PO 2	PO 3	PO 4	PO 5	PSO 1	PSO 3
CO 2	1	3	1		2	2	
CO 3	2		3		1	2	1
CO 4		1	3	2		1	
CO 5	1	3	3				

XII. MAPPING COURSE LEARNING OUTCOMES LEADING TO THE ACHIEVEMENT OF PROGRAM OUTCOMES AND PROGRAM SPECIFIC OUTCOMES:

(CLOs)	Program Outcomes (POs)												Program Specific Outcomes (PSOs)		
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO 2	PSO 3
CLO 1	3												3		1
CLO 2	3			2									1		1
CLO 3	3	3													
CLO 4	3	2											2		
CLO 5		2	3												1
CLO 6	3				1										
CLO 7		2			1								2		
CLO 8		2													
CLO 9	3		3										2		
CLO 10	3												2		
CLO 11			2		1										
CLO 12	3		2										2		
CLO 13			2	2											1
CLO 14			2	2									2		
CLO 15			2												
CLO 16		2	3												
CLO 17		2	3										2		
CLO 18		1	3												
CLO 19			3		2										

(CLOs)	Program Outcomes (POs)												Program Specific Outcomes (PSOs)		
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO 2	PSO 3
CLO 20	3	2													

3 = High; 2 = Medium; 1 = Low

XIII. ASSESSMENT METHODOLOGIES – DIRECT

CIE Exams	PO 1,PO 2, PO3, PO 4, PO5,PSO1, PSO3	SEE Exams	PO 1,PO 2, PO 3, PO 4, PO5 , PSO1, PSO3	Assignments	PO 2, PO 3, PSO1, PSO3	Seminars	PO 4, PO 5, PSO1, PSO3
Laboratory Practices	PO 1, PO5	Student Viva	-	Mini Project	-	Certification	-
Term Paper	-						

XIV. ASSESSMENT METHODOLOGIES - INDIRECT

✔	Early Semester Feedback	✔	End Semester OBE Feedback
✘	Assessment of Mini Projects by Experts		

XV. SYLLABUS

Unit-I	DATA WAREHOUSING
Introduction to Data warehouse, Differences between OLAP and OLTP, A Multi dimensional data model- Star, Snow flake and Fact constellation schemas, Measures, Concept hierarchy, OLAP Operations in the Multidimensional Data Model, Data warehouse architecture- A three tier Data warehouse architecture, Data warehouse Back-End Tools and Utilities, Metadata Repository, types of OLAP servers, Data warehouse Implementation, Data Warehouse models- Enterprise warehouse, Data Marts, Virtual warehouse.	
Unit-II	**DATA MINING**
Introduction, what is Data Mining, Definition, Knowledge Discovery in Data (KDD), Kinds of data bases, Data mining functionalities, Classification of data mining systems, Data mining task primitives, Data Preprocessing: Data cleaning, Data integration and transformation, Data reduction, Data discritization and Concept hierarchy.	
Unit-III	**ASSOCIATION RULE MINING**
Association Rules: Problem Definition, Frequent item set generation, The APRIORI Principle, support and confidence measures, association rule generation; APRIORI algorithm. FP-Growth Algorithms, Compact Representation of Frequent item Set-Maximal Frequent item set, closed frequent item set.	
Unit-IV	**CLASSIFICATION AND PRIDICTION**
Issues Regarding Classification and Prediction, Classification by Decision Tree Induction, Bayesian Classification, Classification by Back propagation, Classification Based on Concepts from Association Rule Mining, Other Classification Methods, Prediction, Classifier Accuracy.	
Unit-V	**CLUSTERING**

Types of data, categorization of major clustering methods, K-means partitioning methods, hierarchical methods, density based methods, grid based methods, model based clustering methods, outlier analysis. Mining Complex Types of Data: Multi dimensional Analysis and Descriptive Mining of Complex, Data Objects, Mining Spatial Databases, Mining Multimedia Databases, Mining Time-Series and Sequence Data, Mining Text Databases, Mining the World Wide Web. S
Text Books:
1. Jiawei Han, Michelin Kamber, "Data Mining-Concepts and techniques", Morgan Kaufmann Publishers, Elsevier, 2nd Edition, 2006 2. Alex Berson, Stephen J.Smith, "Data warehousing Data mining and OLAP", Tata McGraw-Hill, 2nd Edition, 2007
Reference Books:
1. Arum K Pujari, "Data Mining Techniques", 3rd Edition, Universities Press, 2005 2. PualrajPonnaiah, Wiley, "Data Warehousing Fundamentals", Student Edition, 2004. 3. Ralph Kimball, Wiley, "The Data warehouse Life Cycle Toolkit", Student Edition, 2006.

XVI. COURSE PLAN:

The course plan is meant as a guideline. Probably there may be changes.

Lecture No	Topics to be covered	Course Learning Outcomes (CLOs)	Reference
1-3	Introduction to Data warehouse, Difference between operational database systems and data warehouses	CLO 1	T1: 3.1
4-7	Data warehouse Architecture and its components	CLO 2	T1: 3.3
8-10	Multi dimensional data model -Show-Flake Schema, Fact Consultation, Fact Table, Dimension Table	CLO 3	T1: 3.2
11-14	OLAP Cube, OLAP Operations, OLAP Server Architecture-ROLAP, MOLAP and HOLAP.	CLO 4	T1: 3.4-3.5
15-16	Introduction, Fundamentals of Data Mining, Definition, KDD, Challenges, Data Mining Tasks.	CLO 5	T1: 1.1-1.7
17-19	Data Processing, Data Cleaning, , Dimensionality Reduction	CLO 6	T1: 2.1-2.5
20-22	Feature Subset Selections, Data Transformation. Discretization and Measures of Similarity and Dissimilarity-Basics.	CLO 7	T1: 2.3-2.4
23-25	Association Rules: Problem definition, Frequent item set generation,	CLO 8	T1: 5.3
26-27	The APRIORI Principle. Support and confidence measures, association rule generation; APRIORI algorithm.	CLO 9	T1: 5.2
28-29	The partition algorithms, fp-growth Algorithm.	CLO 10	T1: 5.2.2
30	Compact Representation of Frequent item Set-Maximal Frequent item set closed frequent item set.	CLO 11	T1: 5.2.4
31-35	Classification and prediction, basic concepts, decision tree induction,	CLO 12	T1: 6.1-6.2
36-37	Bayesian classification	CLO 13	T1: 6.4
38-39	Rule based classification	CLO 14	T1: 6.5

40-44	Classification by back propagation.	CLO 15	T1: 6.6
40-45	Clustering Analysis	CLO 16	T1: 7.1-7.3
46-47	hierarchical methods,	CLO 17	T1: 7.5
48-52	density based methods	CLO 18	T1: 7.6
53-56	grid based methods, outlier analysis	CLO 19	T1: 7.7
57-60	Mining Complex Types of Data	CLO 20	T1: 7.11

XVII. GAPS IN THE SYLLABUS - TO MEET INDUSTRY / PROFESSION REQUIREMENTS:

S no	Description	Proposed actions	Relevance with POs	Relevance with PSOs
1	Spatial Data mining	Seminars / Guest Lectures	PO 1, PO 2, PO 3	PSO 1
2	Graph mining, social network analysis	Seminars	PO 2, PO 5	PSO 3
3	Mining the world wide web	Seminars NPTEL	PO 1, PO 3, PO 4	PSO 2

Prepared by:
Mr. Ch Suresh Kumar Raju, Assistant Professor

HOD, IT