

Fr. Conceicao Rodrigues College Of Engineering
Department of Artificial Intelligence and Data Science Engineering

T.E. (AI DS) (semester V) (2022-2023)
Course Outcomes & Assessment Plan

Subject: Data Warehousing and Mining (DWM-CSC504)

Credits-3

Course Objectives:

1. To create awareness of how enterprise can organize and analyze large amounts of data by creating a Data Warehouse
2. To introduce the concept of data Mining as an important tool for enterprise data management and as a cutting edge technology for building competitive advantage.
3. To enable students to effectively identify sources of data and process it for data mining
4. To make students well versed in all data mining algorithms, methods of evaluation
5. To impart knowledge of tools used for data mining, and study web mining

Teaching Scheme

Course Code	Course Name	Teaching Scheme			Credits Assigned			
		Theory	Practical	Tutorial	Theory	Practical/Ora l	Tut	Credits
CSC504	Data Warehousing and Mining	03	--	--	03	--	---	03
CSL503	Data Warehousing and Mining Lab	--	02	--	--	1	--	01

Examination Scheme

Course Code	Course Name	Theory Marks				Term Work	Practical & Oral	Total
		Internal Assessment			End Sem Exam			
		Test 1	Test2	Avg				
CSC502	Data Warehousing and Mining	20	20	20	80 (3hr)	--	---	100
CSL502	Data Warehousing and Mining Lab					25	25	50

Syllabus: Prerequisite: Database Management concepts**1. Data Warehouse and OLAP (09)**

Data Warehousing, Dimensional Modeling and OLAP The Need for Data Warehousing; Data Warehouse Defined; Benefits of Data Warehousing ; Features of a Data Warehouse; Data Warehouse Architecture; Data Warehouse and Data Marts; Data Warehousing Design Strategies. Dimensional Model Vs ER Model; The Star Schema, The Snowflake Schema; Fact Tables and Dimension Tables; Factless Fact Table; Updates To Dimension Tables, Primary Keys, Surrogate Keys & Foreign Keys; Aggregate Tables; Fact Constellation Schema or Families of Star Need for Online Analytical Processing; OLTP vs OLAP; OLAP Operations in a cube: Roll-up, Drilldown, Slice, Dice, Pivot ; OLAP Models: MOLAP, ROLAP, HOLAP. Major steps in ETL Process.

1. Introduction to Data Mining ,Data Exploration and Data Preprocessing (08)

Data Mining Task primitives,Architecture,KDD process,Issues in data Mining,Types of Attributes; Statistical Description of Data; Data Visualization; Measuring similarity and dissimilarity. Why Preprocessing? Data Cleaning; Data Integration; Data Reduction: Attribute subset selection, Histograms, Clustering and Sampling; Data Transformation & Data Discretization: Normalization, Binning, Histogram Analysis and Concept hierarchy generation.

2. Classification (06)

Basic Concepts; Classification methods: 1. Decision Tree Induction: Attribute Selection Measures, Tree pruning. 2. Bayesian Classification: Naïve Bayes" Classifier. Prediction: Structure of regression models; Simple linear regression, Multiple linear regression. Accuracy and Error measures, Precision, Recall

3. Clustering (04)

Cluster Analysis: Basic Concepts; Partitioning Methods: K-Means, KMediods; Hierarchical Methods: Agglomerative, Divisive, BIRCH;DensityBased Methods: DBSCAN What are outliers? Types, Challenges; Outlier Detection Methods: Supervised, Semi Supervised, Unsupervised, Proximity based, Clustering Based

4. Frequent Pattern (08)

Market Basket Analysis, Frequent Itemsets, Closed Itemsets, and Association Rules; Frequent Pattern Mining, Efficient and Scalable Frequent Itemset Mining Methods, The Apriori Algorithm for finding Frequent Itemsets Using Candidate Generation, Generating Association Rules from Frequent Itemsets, Improving the Efficiency of Apriori, A pattern growth approach for mining Frequent Itemsets; Mining Frequent itemsets using vertical data formats; Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules; From Association Mining to Correlation Analysis, lift, ; Introduction to Constraint-Based Association Mining

5. Web Mining (04)

Introduction to Web content Mining, Crawlers, Personalization, Webstructure mining, Page rank,, Clever, Web Usage Mining

Internal Assessment:

Assessment consists of two class tests of 20 marks each. The first-class test is to be conducted when

approx. 40% syllabus is completed and second class test when additional 40% syllabus is completed. Duration of each test shall be one hour.

End Semester Theory Examination:

1. Question paper will consist of 6 questions, each carrying 20 marks.
2. The students need to solve a total of 4 questions.
3. Question No.1 will be compulsory and based on the entire syllabus.
4. Remaining question (Q.2 to Q.6) will be selected from all the modules.

Lecture Plan : SEM V-DWM-CSC504**Modes of Content Delivery:**

i	Class Room Teaching	v	Self-Learning Online Resources	ix	Industry Visit
ii	Tutorial	vi	Slides	x	Group Discussion
ii	Remedial Coaching	vii	Simulations/Demonstrations	xi	Seminar
i	Lab Experiment	vii	Expert Lecture	xi	Case Study
v		i		i	

Term : 18th July – 30 Oct 2022**(UT1 : 05 Sept - 07 Sept) (UT2 : 17Oct -19 Oct)**

No.	Portion to be covered	Planned date	Actual date	Content Delivery - Reference /Assessment Method
1	Prerequisite, CO Discussion Introduction to Data Warehouse,	18/07/2022	20/07/2022	PPT/BlackBoard 2
2	Data Warehousing, Dimensional Modeling and OLAP The Need for Data Warehousing; Data Warehouse Defined; Benefits of Data Warehousing ;	22/07/2022	22/07/2022	PPT 2
3	Features of a Data Warehouse; Data Warehouse Architecture; Data Warehouse and Data Marts;	25/07/2022	25/07/2022	PPT/BlackBoard 2
4	Data Warehousing Design Strategies. Dimensional Model Vs ER Model; The Star Schema, The Snowflake Schema; Fact Tables and Dimension Tables;	27/07/2022	27/07/2022	PPT/BlackBoard 2
5	Factless Fact Table; Updates To Dimension Tables, Primary Keys, Surrogate Keys & Foreign Keys; Aggregate Tables; Fact Constellation Schema or Families of Star	29/07/2022	29/07/2022	PPT 2
6	Need for Online Analytical Processing; OLTP vs OLAP; OLAP Operations in a cube:	1/8/2022	3/8/2022	PPT/Lab Demo
7	Roll-up, Drilldown, Slice, Dice, Pivot ;	3/8/2022	5/8/2022	Lab Demo
8	Solved problems on schema design and OLAP operations	5/8/2022	8/8/2022	PPT/BlackBoard
9	OLAP Models: MOLAP, ROLAP, HOLAP. Major steps in ETL Process	8/8/2022	9/8/2022	PPT/BlackBoard
10	Data Mining Task primitives,Architecture,KDD process,Issues in data Mining,	10/8/2022	9/8/2022	PPT/BlackBoard
11	Types of Attributes; Statistical Description of Data;	12/8/2022	10/8/2022	PPT/BlackBoard

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

No.	Portion to be covered	Planned date	Actual date	Content Delivery - Reference /Assessment Method
12	Data Visualization; Measuring similarity and dissimilarity.	17/08/2022	12/8/22	PPT/BlackBoard
13	Why Preprocessing? Data Cleaning;	17/08/2022	12/8/22	PPT/BlackBoard
14	Data Integration; Data Reduction: Attribute subset selection, Histograms, Clustering and Sampling;	22/08/2022	17/08/2022	PPT/BlackBoard
15	Data Transformation & Data Discretization:	24/08/2022	22/08/2022	PPT/Blackboard
16	Normalization, Binning, Histogram Analysis and Concept hierarchy generation.	26/08/2022	24/08/2022 29/08/2022	PPT/BlackBoard
17	Basic Concepts; Classification methods: 1. Decision Tree Induction: Attribute Selection Measures, Prediction:	29/08/2022	29/08/2022 09/09/2022	PPT
18	Decision Tree Induction: Attribute Selection Measures,	09/09/2022	12/09/2022 13/09/2022	PPT/BlackBoard
19	Solved examples based on DT	12/09/2022	14/09/2022	BlackBoard chalk
20	Tree pruning. Bayesian Classification: Naïve Bayes Classifier.	14/09/2022	16/09/2022	PPT/BlackBoard
21	Solved examples based on Bayesian Classification	16/09/2022	19/09/2022	PPT/BlackBoard
22	Structure of regression models; Simple linear regression, Multiple linear regression.	19/09/2022	19/09/2022	PPT/Blackboard
23	Accuracy and Error measures, Precision, Recall	21/09/2022	21/09/2022	PPT/BlackBoard
24	Cluster Analysis: Basic Concepts; Partitioning Methods: K-Means	21/09/2022	21/09/2022	PPT/BlackBoard
25	KMedoids;	22/09/2022	21/09/2022	PPT/BlackBoard
26	Hierarchical Methods: Agglomerative, Divisive,	23/09/2022	22/09/2022	BlackBoard chalk
27	Solved problems on K-Mean and K-medoid	24/09/2022	22/09/2022	BlackBoard chalk
28	BIRCH;DensityBased Methods: DBSCAN	26/09/2022	23/09/2022	PPT/BlackBoard
29	What are outliers? Types, Challenges;	27/09/2022	25/09/2022	Online on Google classroom, PPT

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

No.	Portion to be covered	Planned date	Actual date	Content Delivery - Reference /Assessment Method
30	Outlier Detection Methods: Supervised, Semi Supervised, Unsupervised, Proximity based, Clustering Based	28/09/2022	25/09/2022	Online on Google classroom, PPT
31	Market Basket Analysis, Frequent Itemsets, Closed Itemsets, and Association Rules;	30/09/2022	26/09/2022	BlackBoard chalk
32	Frequent Pattern Mining, Efficient and Scalable Frequent Itemset Mining Methods, The Apriori Algorithm for finding Frequent Itemsets Using Candidate Generation, Generating Association Rules from Frequent Itemsets,	03/10/2022	28/09/2022	BlackBoard chalk
33	Solve problems based on Apriory algorithm	07/10/2022	30/09/2022	BlackBoard and Chalk
34	Improving the Efficiency of Apriori, A pattern growth approach for mining Frequent Itemsets;	08/10/2022	3/10/2022	BlackBoard and Chalk
35	Mining Frequent itemsets using vertical data formats; Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules;	10/10/2022	07/10/2022	BlackBoard and Chalk
36	From Association Mining to Correlation Analysis, lift, ; Introduction to Constraint-Based Association Mining	12/10/2022	10/10/2022	BlackBoard and Chalk
37	Introduction to Web content Mining, Page rank,, Clever, Web Usage Mining	14/10/2022	12/10/2022	BlackBoard and Chalk
38	Crawlers, Personalization,	20/10/2022	14/10/2022	PPT/blackboard
39	Webstructure mining,	21/10/2022	21/10/2022	Online on google meet, PPT
40	Page rank,, Clever, Web Usage Mining	27/10/2022	27/10/2022	Online on google meet, PPT

Total Lectures : 40

Text Books:

1. Paulraj Ponniah, "Data Warehousing: Fundamentals for IT Professionals", Wiley India.
2. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann
3. Reema Theraja, "Data warehousing, Oxford University Press.
4. M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education.

Reference Books:

1. Ian H. Witten, Eibe Frank and Mark A. Hall, "Data Mining ", 3rd Edition Morgan Kaufmann publisher.
2. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Introduction to Data Mining", Person Publisher.

Useful Links:

- 1 www.leetcode.com
- 2 www.hackerrank.com
- 3 www.cs.usfca.edu/~galles/visualization/Algorithms.html
- 4 www.codechef.co

Course Outcomes: [Target 2.6]

After successful completion of the course students will be able to:

CSC504.1 : Organize strategic data in an enterprise and build a data Warehouse.

CSC504.2: Analyze data using OLAP operations so as to take strategic decisions and Demonstrate an understanding of the importance of data mining.

CSC504.3: Organize and Prepare the data needed for data mining using pre preprocessing techniques

CSC504.4: Implement the appropriate data mining methods like classification, clustering or Frequent Pattern mining on large data sets

CSC504.5: Define and apply metrics to measure the performance of various data mining algorithms

CSC504.6: Understand Concepts related to Web mining

Mapping of CO and PO/PSO

Relationship of course outcomes with program outcomes: Indicate 1 (low importance), 2 (Moderate Importance) or 3 (High Importance) in respective mapping cell.

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1
CSC504.1	3	3	3										3
CSC504.2	3	3	2	2	3							2	3
CSC504.3	3	3			3							3	2

CSC504.4	3	3	2	2	3							2		2
CSC504.5	3	3	2	2	3									3
CSC504.6	3	2	2	2										2
TOTAL	18	17	11	8	12	0	0	0	6	6	0	7		15
CO-PO MATRIX	3	2.83	1.83	1.33	2	0	0	0	1	1	0	1.16		2.5

CO ASSESSMENT TOOLS

	Direct Methods (80%)					Indirect Methods (20%)
CSC504.1	Test 1 (30%)	Lab 7 (20%)	UE -TH (20%)	UE-O (20%)	Assign 1 (10%)	(100%)
CSC504.2	Test1 (30%)	Lab 4 (30%)	UE -TH (20%)	UE-O (10%)	Assign 1 (10%)	(100%)
CSC504.3	Test1 (30%)	Tut (30%)	UE -TH (20%)	UE-O (10%)	Assign 2 (10%)	(100%)
CSC504.4	Test2 (30%)	Lab 6 (30%)	UE -TH (20%)	UE-O (20%)		(100%)
CSC504.5	Test2 (30%)	MP (20%)	UE -TH (20%)	UE-O (20%)	Assign 3 (10%)	(100%)
CSC604.6			UE -TH (50%)	UE-O (50%)		(100%)

Gurriculum Gap/Content Beyond Syllabus:

Sr.No	Gap/Content Beyond Syllabus	Activity	Topic
1	Information Package diagram	Extra Lecture	What is IPD? Contents of IPD
2	Outlier Analysis	Extra Lecture	Outlier Detection Methods: Supervised, Semi Supervised, Unsupervised, Proximity based, Clustering Based
3	Tree Pruning	Extra Lecture	What is tree pruning, importance of tree pruning, tree pruning methods.

4	OLAP operation implementation in Excel	DEMO	OLAP operations implementation using pivot table and visualization using pivot chart, data analysis using Excel pivot table.
---	--	------	--

Rubrics for Assignments

Class : T.E. AI & DS

Semester : V

Assignment No:	
Title:	
Date of Performance:	
Roll No:	
Name of the Student:	

Evaluation:

Indicator	Poor	Average	Good
Timeliness <ul style="list-style-type: none"> Maintains assignment deadline (2) 	Assignment not done (0)	One or More than One week late (1)	Maintains deadline (2)
Completeness and neatness <ul style="list-style-type: none"> Complete all parts of assignment(3) 	N/A	< 80% complete (1-2)	100% complete (3)
Originality <ul style="list-style-type: none"> Extent of plagiarism(2) 	Copied it from someone else(0)	Atleast few questions have been done without copying(1)	Assignment has been solved completely without copying (2)
Knowledge <ul style="list-style-type: none"> In depth knowledge of the assignment(3) 	Unable to answer all questions(0)	Unable to answer some questions (1 or 2)	Able to answer all questions (3)

Signature

Department of AI & DS Engineering**Rubrics for Lab Experiment**

Class : T.E. AI and DS
Semester : V

Subject Name :DWM
Subject Code :CSC504

Practical No:	
Title:	
Date of Performance:	
Roll No:	
Name of the Student:	

Rubrics for assessment of Experiment:

Sr. No	Parameters	Exceed Expectations(EE)	Meet Expectations (ME)	Below Expectations (BE)
1	Timeline (2)	Early or on time (2)	One session late (1)	More than one session late (0)
2	Preparedness (2)	Knows the basic theory related to the experiment very well. (2)	Managed to explain the theory related to the experiment. (1)	Not aware of the theory to the point. (1)
3	Effort (3)	Done expt on their own. (3)	Done expt with help from other. (2)	Just managed. (1)
4	Documentation(2)	Lab experiment is documented in proper format and maintained neatly. (2)	Documented in proper format but some formatting guidelines are missed. (1)	Experiments not written in proper format (0.5)

5	Result (1)	Specific conclusion.(1)	Partially specific conclusion. (0.5)	Not specific at all. (0)
---	------------	-------------------------	---	--------------------------

Assessment Marks:

Timeline(2)	Preparedness(2)	Effort(3)	Documentation(2)	Result(1)	Total(10)

Signature