

Course Code	Course Name	Credit
CSC504	Data Warehousing and Mining	03

Pre-requisite: Database Management concepts

Course Objectives: The course aims:

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

Course Outcomes:

1	Organize strategic data in an enterprise and build a data Warehouse.
2	Analyze data using OLAP operations so as to take strategic decisions and Demonstrate an understanding of the importance of data mining.
3	Organize and Prepare the data needed for data mining using pre preprocessing techniques
4	Implement the appropriate data mining methods like classification, clustering or Frequent Pattern mining on large data sets.
5	Define and apply metrics to measure the performance of various data mining algorithms
6	Understand Concepts related to Web mining

Module		Detailed Content	Hours
1		Data Warehouse and OLAP	
		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	9
2		Introduction to Data Mining ,Data Exploration and Data Preprocessing	8

		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.	
3		Classification	6
		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	
4		Clustering	4
		Cluster Analysis: Basic Concepts; Partitioning Methods: K-Means, KMedoids; Hierarchical Methods: Agglomerative, Divisive, BIRCH;Density-Based Methods: DBSCAN What are outliers? Types, Challenges; Outlier Detection Methods: Supervised, Semi Supervised, Unsupervised, Proximity based, Clustering Based	
5		Frequent Pattern	8
		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	
6		Web Mining	4
		Introduction to Web content Mining, Crawlers, Personalization, Webstructure mining, Page rank,, Clever, Web Usage Mining	

Textbooks:

1	Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition
2	P. N. Tan, M. Steinbach, Vipin Kumar, "Introduction to Data Mining", Pearson Education.
3	Paulraj Ponniah, "Data Warehousing: Fundamentals for IT Professionals", Wiley India.
4	Raghu Ramakrishnan and Johannes Gehrke, "Database Management Systems" 3rd Edition - McGraw Hill
5	Elmasri and Navathe, "Fundamentals of Database Systems", 6th Edition, PEARSON Education

References:

1	Theraja Reema, "Data Warehousing", Oxford University Press, 2009
2	Ralph Kimball, Margy Ross, "The Data Warehouse Toolkit: The Definitive Guide To Dimensional Modeling", 3rd Edition. Wiley India.

3	Michael Berry and Gordon Linoff “Mastering Data Mining- Art & science of CRM”, Wiley Student Edition
4	Michael Berry and Gordon Linoff “Data Mining Techniques”, 2nd Edition Wiley Publications

Assessment:	
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.

Useful Links	
1	https://www.coursera.org/learn/data-warehousing-business-intelligence
2	https://www.coursera.org/specializations/data-mining-foundations-practice
3	https://onlinecourses.nptel.ac.in/noc20_cs12/preview
4	https://nptel.ac.in/courses/106105174