# Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(Autonomous Institute Affiliated to University of Mumbai)

| Course Code | Course Name | Teaching Scheme (Hrs/week) | | | Credits Assigned | | | |
|---|---|---|---|---|---|---|---|---|
| | | L | T | P | L | T | P | Total |
| CPC801 | Data Warehouse and Mining | 4 | - | -- | 4 | - | -- | 4 |
| | | Examination Scheme | | | | | | |
| | | ISE | | MSE | | ESE | | |
| | | 10 | | 30 | | 100 (60% Weightage) | | |

| Pre-requisite Course Codes | | CSC404 (Database Management System)<br>CPC603 (Distributed Database) |
|---|---|---|
| At end of successful completion of this course, student will be able to | | |
| Course Outcomes | CO1 | Discuss the need of data warehouse and the concepts of data warehousing. |
| | CO2 | Describe the ETL process and illustrate the OLAP operations |
| | CO3 | Express the concepts of data mining, data exploration , preprocessing |
| | CO4 | Apply algorithms in data mining and data warehousing; assess the strengths and weaknesses of the algorithms, identify the application area of algorithms |

| Module No. | Topics | Ref. | Hrs. |
|---|---|---|---|
| 1 | **Introduction to Data Warehousing**<br>The Need for Data Warehousing; Increasing Demand for StrategicInformation; Inability of Past Decision Support System; Operational V/sDecisional Support System; Data Warehouse Defined;Benefits of DataWarehousing;Features of a Data Warehouse;TheInformationFlowMechanism;Role of Metadata; Classification of Metadata; Data WarehouseArchitecture; Different Types of Architecture;Data Warehouse and DataMarts; Data Warehousing Design Strategies. | 1,3 | **04** |
| 2 | **Dimensional Modeling**<br>Data Warehouse Modeling Vs Operational Database Modeling; DimensionalModel Vs ER Model; Features of a Good Dimensional Model; The StarSchema; How Does a Query Execute? The Snowflake Schema; Fact Tablesand Dimension Tables; The Factless Fact Table; Updates To DimensionTables: Slowly Changing Dimensions, Type 1 Changes, Type 2 Changes,Type 3 Changes, Large Dimension Tables, Rapidly Changing or LargeSlowly Changing Dimensions, Junk Dimensions, Keys in the DataWarehouse Schema, Primary Keys, Surrogate Keys & Foreign Keys;Aggregate Tables; Fact Constellation Schema or Families of Star. | 1,3 | **06** |
| 3 | **ETL Process**<br>Challenges in ETL Functions; Data Extraction; Identification of | 1,2,3 | **06** |

| | | | |
|---|---|---|---|
| | DataSources; Extracting Data: Immediate Data Extraction, Deferred DataExtraction; Data Transformation: Tasks Involved in Data Transformation,Data Loading: Techniques of Data Loading, Loading the Fact Tables andDimension Tables Data Quality; Issues in Data Cleansing. | | |
| **4** | **Online Analytical Processing (OLAP)**<br>Need for Online Analytical Processing; OLTP V/s OLAP; OLAP and Multidimensional Analysis; Hypercubes; OLAP Operations inMultidimensional Data Model;OLAP Models: MOLAP, ROLAP, HOLAP,DOLAP; | 1,3,6,9 | **04** |
| **5** | **Introduction to data mining**<br>What is Data Mining; Knowledge Discovery in Database (KDD), What canbe Data to be Mined, Related Concept to Data Mining, Data MiningTechnique, Application and Issues in Data Mining | 1,3,4,5 | **02** |
| **6** | **Data Exploration**<br>Types of Attributes; Statistical Description of Data; Data Visualization;Measuring similarity and dissimilarity. | 1,7 | **02** |
| **7** | **Data Preprocessing**<br>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. | 1,8 | **04** |
| **8** | **Classification**<br>Basic Concepts; Classification methods:Decision Tree Induction: Attribute Selection Measures, Tree pruning, Bayesian Classification: Naïve Bayes' Classifier, Prediction: Structure of regression models; Simple linear regression, Multiple linear regression, Model Evaluation & Selection: Accuracy and Error measures, Holdout,Random Sampling, Cross Validation, Bootstrap; Comparing Classifierperformance using ROC Curves, Combining Classifiers: Bagging, Boosting, Random Forests. | 1,4,8 | **06** |
| **9** | **Clustering**<br>What is clustering? Types of data, Partitioning Methods (K-Means, K-Medoids) Hierarchical Methods(Agglomerative, Divisive, BRICH),Density-Based Methods ( DBSCAN, OPTICS) | 1,4,8 | **06** |
| **10** | **Mining Frequent Pattern and Association Rule**<br>Market Basket Analysis, Frequent Itemsets, Closed Itemsets, andAssociation Rules; Frequent Pattern Mining, Efficient and Scalable FrequentItemset Mining Methods, The Apriori Algorithm for finding FrequentItemsets Using Candidate Generation, Generating Association Rules fromFrequentItemsets, Improving the Efficiency of Apriori, A pattern growthapproach for mining Frequent Itemsets; Mining Frequent itemsetsusingvertical data formats; Mining | 1,4 | **08** |

| | | | |
|---|---|---|---|
| | closed and maximal patterns; Introduction toMiningMultilevel Association Rules andMultidimensional associationRules; From Association Mining to Correlation Analysis, Pattern EvaluationMeasures; Introduction to Constraint-Based Association Mining. | | |
| | | **Total** | **48** |

**References:**


[1] Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition
[2] PaulrajPonniah, "Data Warehousing: Fundamentals for IT Professionals", Wiley India
[3] ReemaTheraja "Data warehousing", Oxford University Press.
[4] M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education
[5] Randall Matignon, "Data Mining using SAS enterprise miner ", Wiley Student edition.
[6] Alex Berson , S. J. Smith, "Data Warehousing, Data Mining & OLAP" , McGraw Hill.
[7] VikramPudi&Radha Krishna, "Data Mining", Oxford Higher Education.
[8] Daniel Larose, "Data Mining Methods and Models", Wiley India.
[9] P.S.Deshpande, "*SQL & PL/SQL for Oracle 11 g*", dreamtech PRESS.