RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based on AICTE Flexible Curricula

CSE-Data Science/Data Science, V semester

CD503 (A) Data Mining & Warehousing

COURSE OBJECTIVES:

Student should understand the value of Historical data and data mining in solving real-world problems. Student should become affluent with the basic Supervised and unsupervised learning algorithms commonly used in data mining.
Student develops the skill in using data mining for solving real-world problems.

Unit 1: Data Warehousing: Introduction, Delivery Process, Data warehouse Architecture, Data Preprocessing: Data cleaning, Data Integration and transformation, Data reduction. Data warehouse Design: Dataware house schema, Partitioning strategy Data warehouse Implementation, Data Marts, Meta Data, Example of a Multidimensional Data model, Introduction to Pattern Warehousing.

Unit 2: OLAP Systems: Basic concepts, OLAP queries, Types of OLAP servers, OLAP operations etc. Data Warehouse Hardware and Operational Design: Security, Backup And Recovery,

Unit 3: Introduction to Data & Data Mining: Data Types, Quality of data, Data Preprocessing, Similarity measures, Summary statistics, Data distributions, Basic data mining tasks, Data Mining V/s knowledge discovery in databases. Issues in Data mining, Introduction to Fuzzy sets and fuzzy logic.

Unit 4: Supervised Learning (Classification): Statistical-based algorithms, Distance-based algorithms, Decision tree-based algorithms, Neural network-based algorithms, Rule-based algorithms, Probabilistic Classifiers

Unit 5: Clustering & Association Rule mining: Hierarchical algorithms, Partitional algorithms, Clustering large databases – BIRCH, DBSCAN, CURE algorithms. Association rules : Parallel and distributed algorithms such as Apriori and FP growth algorithms.

Books Recommended:

Text Books:

1. Pang – ningTan , Steinbach & Kumar, “Introduction to Data Mining”, Pearson Edu, 2019.
2. Jaiwei Han, MichelineKamber, “Data Mining : Concepts and Techniques”, Morgan Kaufmann Publishers.

Reference Books:

1. Margaret H. Dunham, “Data Mining : Introductory and Advanced topics”, Pearson Edu., 2009.
2. Anahory& Murray, “Data Warehousing in the Real World”, Pearson Edu., 2009.

COURSE OUTCOMES

After completion of this course, the students would be able to:

- CO1. Understand the need of designing Enterprise data warehouses and will be enabled to approach business problems analytically by identifying opportunities to derive business.
- CO2. Compare and contrast various methods for storing & retrieving data from different data sources/repository.
- CO3. Ascertain the application of data mining in various areas and Preprocess the given data and visualize it for a given application or data exploration/mining task
- CO4. Apply supervised learning methods to given data sets such as classification and its various types.
- CO5. Apply Unsupervised learning methods to given data sets such as clustering and its various types.
- CO6. Apply Association rule Mining to various domains.

RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based on AICTE Flexible Curricula

CSE-Data Science/Data Science, V semester

CD503 (B) PATTERN RECOGNITION

Unit-I

Introduction – Definitions, datasets for Pattern, Application Areas and Examples of pattern recognition, Design principles of pattern recognition system, Classification and clustering, supervised Learning, unsupervised learning and adaptation, Pattern recognition approaches, Decision Boundaries, Decision region , Metric spaces, distances.

Unit -II

Classification: introduction, application of classification, types of classification, decision tree, naïve bayes, logistic regression , support vector machine, random forest, K Nearest Neighbour Classifier and variants, Efficient algorithms for nearest neighbour classification, Different Approaches to Prototype Selection, Combination of Classifiers, Training set, test set, standardization and normalization.

Unit – III

Different Paradigms of Pattern Recognition, Representations of Patterns and Classes, Unsupervised Learning & Clustering: Criterion functions for clustering, Clustering Techniques: Iterative square -error partitional clustering – K means, hierarchical clustering, Cluster validation.

Unit -IV

Introduction of feature extraction and feature selection, types of feature extraction , Problem statement and Uses, Algorithms - Branch and bound algorithm, sequential forward / backward selection algorithms, (l,r) algorithm.

Unit -V

Recent advances in Pattern Recognition, Structural PR, SVMs, FCM, Soft computing and Neuro-fuzzy techniques, and real-life examples, Histograms rules, Density Estimation, Nearest Neighbor Rule, Fuzzy classification.

REFERENCE BOOKS:

1. Richard O. Duda, Peter E. Hart and David G. Stork, "Pattern Classification", 2nd Edition, John Wiley, 2006.
2. C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2009.
3. S. Theodoridis and K. Koutroumbas, "Pattern Recognition", 4th Edition, academic Press, 2009.
4. Robert Schalkoff, "Pattern Recognition: statistical, structural and neural approaches", JohnWiley&sons ,Inc, 2007.

RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based on AICTE Flexible Curricula

CSE-Data Science/Data Science, V semester

CD503 (C) Introduction to Toolkits for Data Science

Course Content:

Unit 1: Python for Data Science: Review of Numpy, Pandas and Scikit-learn.Supervised Learning Techniques packages/toolkit for regression and classification: - Decision Trees, Naive Bayes, Classification, Support vector machines, Random Forest, Neural network, Ensemble Methods, Ordinary Least Squares Regression, Logistic Regression, etc. Unsupervised Learning, Clustering: k-means, adaptive hierarchical clustering, Gaussian mixture, Optimization Using Evolutionary Techniques etc.

Unit 2: R for Data Science: Basic of R and RStudio. R data structures: vectors, factors, lists, arrays, matrices, and data frames. Working with data: Import data into R and visualize data. Data Analytics Software: Weka, Orange, Rapidminer, Minitab, PowerBI, GitHub, Google Colab.

Unit 3. Introduction to Deep Learning: Basics of TensorFlow and keras,Basics of PyTorch, perform style transfer of one image to another, Perform text generation, and sentiment analysis with PyTorch. Neural networks that recognize objects, improve the accuracy of object recognition using CNN, use pre-trained models to build state-of-the-art classifiers, Saving and Loading models, Time series forecasting with RNNs, and LSTMs,

Unit 4: Introduction to Time Series Analysis: Time series regression and exploratory data analysis toolkits: ARMA/ARIMA models, model identification/estimation/linear operators, Fourier analysis, spectral estimation, and state-space models.

Unit 5: Cloud Computing for Data Science: Implementation of Machine Learning and Deep learning through AWS/Azure platform.

Version controlling tools for data science projects. Case studies of data science projects.

Books Recommended:

- Brockwell& Davis (2016) Introduction to Time Series and Forecasting, 3rd edition, Springer
- Cryer& Chan (2008) Time-Series Analysis with Applications in R, Springer
- Prado & West (2010) Time Series: Modeling, Computation, and Inference Chapman & Hall
- Petris, Petrone, Campagnoli (2009) Dynamic Linear Models with R, Springer
- Ruppert& Matteson (2016) Statistics and Data Analysis for Financial Engineering with R examples, 2nd Edition, Springer
- R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, 1st Edition, O'reilly publication.