II YEAR III SEMESTER PAPER– III

DATA MINING AND DATA ANALYSIS

Objective

- To learn data analysis techniques.
- To understand Data mining techniques and algorithms.
- Comprehend the data mining environments and application.

Outcome

Students who complete this course will be able to

1. To understand and demonstrate data mining
2. Compare various conceptions of data mining as evidenced in both research and application.
3. Characterize various kinds of patterns that can be discovered by association rule mining.
4. Evaluate mathematical methods underlying the effective application of data mining.
5. To Analyze the data using statistical methods
6. Gain hands-on skills and experience on data mining tools.

Unit-I

Data mining - KDD Vs Data Mining, Stages of the Data Mining Process-Task Primitives, Data Mining Techniques – Data Mining Knowledge Representation. Major Issues in Data Mining – Measurement and Data – Data Preprocessing – Data Cleaning - Data transformation- Feature Selection - Dimensionality reduction

Unit-II: Predictive Analytics

Classification and Prediction **-** Basic Concepts of Classification and Prediction, General Approach to solving a classification problem- Logistic Regression - LDA - Decision Trees: Tree Construction Principle – Feature Selection measure – Tree Pruning - Decision Tree construction Algorithm, Random Forest, Bayesian Classification-Accuracy and Error Measures- Evaluating the Accuracy of the classifier / predictor- Ensemble methods and Model selection.

Unit-III : Classification and Descriptive Analytics

Rule Based Classification – Classification by Back propagation – Support Vector Machines – Associative Classification – Lazy Learners – Other Classification Methods – Prediction. Descriptive Analytics - Mining Frequent Itemsets - Market based model – Association and Sequential Rule Mining

Unit - IV : Cluster Analysis

Cluster Analysis: Basic concepts and Methods – Cluster Analysis – Partitioning methods –

Hierarchical methods – Density Based Methods – Grid Based Methods – Evaluation of Clustering – Advanced Cluster Analysis: Probabilistic model based clustering – Clustering High – Dimensional Data – Clustering Graph and Network Data – Clustering with Constraints- Outlier Analysis.

Unit-V: Factor Analysis

Factor Analysis: Meaning, objectives and Assumptions, Designing a factor analysis, Deriving factors and assessing overall factors, Interpreting the factors and validation of factor analysis.

References

1. Adelchi Azzalini, Bruno Scapa, "Data Analysis and Data mining" , 2nd Ediiton, Oxford Univeristy Press Inc., 2012.
2. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2011.
3. Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining & OLAP", 10th
   Edition, TataMc Graw Hill Edition , 2007.
4. G.K. Gupta, "Introduction to Data Mining with Case Studies", 1st Edition, Easter Economy Edition, PHI, 2006.
5. Joseph F Hair, William C Black etal, "Multivariate Data Analysis", Pearson Education,
   7th edition, 2013.

Student Activity

Case Study I: Analysis and Forecasting of House Price Indices

Case Study II: Customer Response Prediction and Profit Optimization

Case Study III: Iris Species Prediction

RECOMMENDED CO-CURRICULAR ACTIVITIES:

(Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)

2. Student seminars (on topics of the syllabus and related aspects (individual activity))

3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))

4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity

B. General

1. Group Discussion

2. Try to solve MCQ’s available online.

3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)

2. Closed-book and open-book tests

3. Problem-solving exercises

4. Practical assignments and laboratory reports

5. Observation of practical skills

6. Individual and group project reports like “Movie Lens Data Analysis”, "COVID-19 Analysis", etc.

7. Efficient delivery using seminar presentations,

8. Viva voce interviews.

9. Computerized adaptive testing, literature surveys and evaluations,

10. Peers and self-assessment, outputs form individual and collaborative work

II YEAR III SEMESTER PAPER– III

DATA MINIG AND DATA ANALYSIS LAB

1. Data Analysis – Getting to know the Data (Using ORANGE WEKA or R Programming)
    - Parametric – Means, T-Test, Correlation
    - Prediction for numerical outcomes – Linear regression, Multiple Linear Regression
    - Correlation analysis
    - Preparing data for analysis
        - Pre-Processing techniques

2. Data Mining (Using ORANGE WEKA or R Programming)
    - Implement clustering algorithm
    - Implement Association Rule mining
    - Implement classification using
        - Decision tree
        - Back Propagation
        - Logistic Regression
        - Decision Tree
        - Random Forest
        - Naive Bayes
        - Support Vector Machines
    - Visualization methods