



SRI VENKATESWARA UNIVERSITY

II Year Syllabus for 4-Year UG Honours in B.Sc. (Data Science) as Major in consonance with Curriculum framework w.e.f. AY 2026-27

Year	Semester	Course	Title of the Course	No. of Hrs /Week	No. of Credits
II	IV	8	Data Mining	3	3
			Data Mining Lab	2	1
		9	Python for Data Analysis and Visualization	3	3
			Python for Data Analysis and Visualization lab	2	1
		10	Document Oriented Database	3	3
			Document oriented Database lab	2	1

SEMESTER-IV
COURSE 8: DATA MINING

Theory Credits: 3 3 hrs/week

Course Objectives:

- Provide an understanding of data warehousing concepts, architecture, and OLAP operations for effective storage, modeling, and analysis.
- Develop knowledge of data mining fundamentals, tasks, and preprocessing techniques to prepare data for mining.
- Introduce students to association rule mining algorithms for discovering hidden patterns and relationships in large datasets.
- Enable learners to apply classification techniques (decision trees, Bayesian, nearest neighbor, rule-based) for predictive modeling.
- Equip students with knowledge of clustering paradigms and algorithms (partitioning, hierarchical, density-based, categorical) for data grouping and pattern discovery.

Course Outcomes:

Upon successful completion of the course, the student will be able to:

1. Explain the architecture, schemas, and OLAP operations of data warehousing and distinguish it from traditional database systems.
2. Apply preprocessing techniques (data cleaning, dimensionality reduction, feature selection, transformation, and similarity measures) to prepare raw data for analysis.
3. Implement various association rule mining algorithms (Apriori, Partition, FP-Growth, etc.) to uncover meaningful relationships within large datasets.
4. Build and evaluate classification models using decision tree algorithms (ID3, C4.5, CART), rule-based classifiers, Bayesian classifiers, and nearest-neighbor methods.
5. Analyze and implement clustering techniques such as K-Means, K-Medoid, DBSCAN, BIRCH, and categorical clustering methods (STIRR, ROCK, CACTUS) for grouping and pattern discovery in different types of datasets.

Unit-1: Data Warehousing:

Introduction to Data Ware House, Differences between Database systems and Data Ware House, Data Ware House characteristics, Data Ware House Architecture and its components, Data Modeling, Schema Design, star and snow-Flake Schema, Fact Constellation, Fact Table, OLAP cube, OLAP Operations.

Unit-2: Data Mining:

What is Data Mining? Data Mining: Definitions, KDD vs Data Mining, Data Mining Tasks, Data Preprocessing- Data Cleaning, Missing Data, Dimensionality Reduction, Feature Subset Selection, Discretization and Binarization, Data Transformation; Measures of similarity and Dissimilarity-Basics. Issues and Challenges in DM, DM Applications- Case Studies

Unit-3: Association Analysis:

Association Rules: What is an Association Rule?, Methods to Discover Association Rules, A Priori Algorithm, Partition Algorithm, Pincer-Search Algorithm, Dynamic Itemset Counting Algorithms, FP-Tree Growth Algorithm, Generalized Association Rule, Association Rules with Item Constraints

Unit-4: Classification:

Definition, What is Decision Tree?, Tree Construction Principle, Best Split, Splitting Indices, Splitting Criteria, Decision Tree Construction Algorithms: CART, ID3, C4.5, Method for Comparing Classifiers, Rule Based Classifiers, Nearest Neighbor Classifiers, Bayesian Classifiers.

Unit-5: Clustering Techniques:

Clustering Paradigms, Partitioning Algorithms (K-Means), k-Medoid Algorithms, Hierarchical Clustering: DBSCAN, BIRCH, Categorical Clustering Algorithms: STIRR, ROCK, CACTUS

Textbooks:

1. Data Mining Techniques, Arun K Pujari, 3rd Edition, Universities Press
2. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, 3rd Edition, Morgan Kaufmann Publishers

Reference Books:

1. K.P. Soman , Shyam Diwakar, V.Ajay ,2006, Insight into Data Mining Theory and Practice, Prentice Hall of India Pvt. Ltd - New Delhi.
2. Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, 2nd edition,

Activities:

Outcome: Explain the architecture, schemas, and OLAP operations of data warehousing and distinguish it from traditional database systems.

Activity: Students will design a **conceptual schema** for a data warehouse using **star or snowflake schema** for a given business case (e.g., sales, hospital, or university records). They will also **demonstrate OLAP operations** (roll-up, drill-down, slice, dice) on a sample dataset using a spreadsheet or OLAP tool.

Evaluation Method: Assessment will be based on correctness and completeness of schema design, appropriateness of fact/dimension tables, and accuracy of OLAP operation outputs. A short viva/quiz will be used to test conceptual understanding.

Outcome 2: Apply preprocessing techniques (data cleaning, dimensionality reduction, feature selection, transformation, and similarity measures) to prepare raw data for analysis.

Activity: Students will be given a **real-world noisy dataset** (with missing values, outliers, redundant features). They will perform:

- Data cleaning (handling missing values, outliers)
- Dimensionality reduction (e.g., PCA)
- Feature selection (e.g., filter/wrapper methods)
- Similarity/dissimilarity measure calculations.

Evaluation Method: Evaluation will consider correctness of preprocessing steps, justification of chosen methods, and clarity of intermediate outputs. Students will submit a short report with before/after comparisons.

Outcome 3: Implement various association rule mining algorithms (Apriori, Partition, FP- Growth, etc.) to uncover meaningful relationships within large datasets.

Activity: Using a **transaction dataset** (e.g., market basket data), students will implement Apriori/FP-Growth in **Python (mlxtend library) or Weka**. They will generate frequent itemsets, derive association rules, and interpret them under given **support, confidence, and lift thresholds**.

Evaluation Method: Submissions will be graded on correctness of generated rules, clarity of code/parameters, quality of interpretation, and ability to link discovered rules to practical business insights.

Outcome 4: Build and evaluate classification models using decision tree algorithms (ID3, C4.5, CART), rule-based classifiers, Bayesian classifiers, and nearest-neighbor methods. **Activity:**

Students will implement at least **two classification algorithms** (e.g., Decision Tree + Naïve Bayes or KNN) on a dataset like **Iris, Titanic, or Student Performance**. They will evaluate models using **confusion matrix, accuracy, precision, recall, and F1-score** and compare results.

Evaluation Method: Assessment will consider correctness of implementation, clarity in performance comparison, and visualization of decision trees/rules. Students must explain why certain models performed better.

Outcome: Analyze and implement clustering techniques such as K-Means, K-Medoid, DBSCAN, BIRCH, and categorical clustering methods (STIRR, ROCK, CACTUS).

Activity: Students will implement at least two clustering algorithms (e.g., K-Means and DBSCAN or BIRCH) on real datasets (e.g., customer segmentation, text data, student groups). They will compare clusters using Silhouette Score, Davies-Bouldin Index, and visualize results using scatter plots/dendrograms.

Evaluation Method: Evaluation will be based on accuracy of clustering implementation, choice of parameters (e.g., k in K-Means, eps in DBSCAN), visualization quality, and clarity in interpretation of clusters.

SEMESTER-IV
COURSE 7: DATA MINING

Practical Credits: 1 2 hrs/week

List of Experiments

Recommended datasets: **weather.arff**, **iris.arff**, **supermarket.arff**, **vote.arff**, **contact-lenses.arff**, or custom CSV datasets.

1. Load datasets in WEKA and explore data formats (ARFF/CSV)
2. Perform data cleaning and handle missing values using filters
3. Apply normalization and discretization on numeric attributes
4. Reduce data using attribute selection and PCA
5. Summarize and visualize data using statistical tools and class-wise comparison in WEKA.
6. Generate association rules using the Apriori algorithm
7. Apply multilevel association rule mining using hierarchical attributes
8. Apply K-means clustering and interpret the cluster outputs.
9. Perform hierarchical clustering and visualize results using dendrograms.
10. Apply Expectation-Maximization (EM) clustering and analyze cluster summaries.
11. Build a decision tree classifier using J48 and evaluate its performance.
12. Perform Naive Bayes classification and compare with decision tree results.
13. Apply rule-based classification using PART or JRip algorithms.
14. Compare classifiers using confusion matrix, accuracy, and ROC curves.
15. Perform basic text preprocessing and clustering using TF-IDF and K-means.

SRI VENKATESWARA UNIVERSITY::TIRUPATI
B.Sc. Data Science Honours
II year IV Semester
COURSE 8: DATA MINING

MODEL QUESTION PAPER

Time: **3 Hours**

Max. Marks: **75**

PART-A

Answer any **FIVE** of the following. Each Question Carries 3 marks.

(5X3=15)

1. Short answer question from unit-I
2. Short answer question from unit-I
3. Short answer question from unit-II
4. Short answer question from unit-II
5. Short answer question from unit-III
6. Short answer question from unit-III
7. Short answer question from unit-IV
8. Short answer question from unit-IV
9. Short answer question from unit-V
10. Short answer question from unit-V

PART-B

Answer **FIVE** of the following. Each Question Carries 12 marks.

(5X12=60)

11. a) Essay question from Unit-I
OR
b) Essay question from Unit-I
12. a) Essay question from Unit-II
OR
b) Essay question from Unit-II
13. a) Essay question from Unit-III
OR
b) Essay question from Unit-III
14. a) Essay question from Unit-IV
OR
b) Essay question from Unit-IV
15. a) Essay question from Unit-V
OR
b) Essay question from Unit-V