JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA
KAKINADA – 533 003, Andhra Pradesh, India

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

III Year – I Semester		**L**	**T**	**P**	**C**
		0	**0**	**3**	**1.5**
DATA WAREHOUSING AND DATA MINING LAB					

Course Objectives: The main objective of the course is to

- Inculcate Conceptual, Logical, and Physical design of Data Warehouses OLAP applications and OLAP deployment
- Design a data warehouse or data mart to present information needed by management in a form that is usable
- Emphasize hands-on experience working with all real data sets.
- Test real data sets using popular data mining tools such as WEKA, Python Libraries
- Develop ability to design various algorithms based on data mining tools.

Course Outcomes: By the end of the course student will be able to

- Design a data mart or data warehouse for any organization
- Extract knowledge using data mining techniques and enlist various algorithms used in information analysis of Data Mining Techniques
- Demonstrate the working of algorithms for data mining tasks such as association rule mining, classification for realistic data
- Implement and Analyze on knowledge flow application on data sets and Apply the suitable visualization techniques to output analytical results

Software Requirements: WEKA Tool/Python/R-Tool/Rapid Tool/Oracle Data mining

List of Experiments:

1. Creation of a Data Warehouse.
   - ➢ Build Data Warehouse/Data Mart (using open source tools like Pentaho Data Integration Tool, Pentaho Business Analytics; or other data warehouse tools like Microsoft-SSIS, Informatica, Business Objects,etc.,)
   - ➢ Design multi-dimensional data models namely Star, Snowflake and Fact Constellation schemas for any one enterprise (ex. Banking, Insurance, Finance, Healthcare, manufacturing, Automobiles, sales etc).
   - ➢ Write ETL scripts and implement using data warehouse tools.
   - ➢ Perform Various OLAP operations such slice, dice, roll up, drill up and pivot

2. Explore machine learning tool “WEKA”
   - ➢ Explore WEKA Data Mining/Machine Learning Toolkit.
   - ➢ Downloading and/or installation of WEKA data mining toolkit.
   - ➢ Understand the features of WEKA toolkit such as Explorer, Knowledge Flow interface, Experimenter, command-line interface.
   - ➢ Navigate the options available in the WEKA (ex. Select attributes panel, Preprocess panel, Classify panel, Cluster panel, Associate panel and Visualize panel)
   - ➢ Study the arff file format Explore the available data sets in WEKA. Load a data set (ex. Weather dataset, Iris dataset, etc.)
   - ➢ Load each dataset and observe the following:
     1. List the attribute names and they types
     2. Number of records in each dataset
     3. Identify the class attribute (if any)
     4. Plot Histogram
     5. Determine the number of records for each class.
     6. Visualize the data in various dimensions

3. Perform data preprocessing tasks and Demonstrate performing association rule mining on data sets
   - Explore various options available in Weka for preprocessing data and apply Unsupervised filters like Discretization, Resample filter, etc. on each dataset
   - Load weather. nominal, Iris, Glass datasets into Weka and run Apriori Algorithm with different support and confidence values.
   - Study the rules generated. Apply different discretization filters on numerical attributes and run the Apriori association rule algorithm. Study the rules generated.
   - Derive interesting insights and observe the effect of discretization in the rule generation process.

4. Demonstrate performing classification on data sets
   - Load each dataset into Weka and run 1d3, J48 classification algorithm. Study the classifier output. Compute entropy values, Kappa statistic.
   - Extract if-then rules from the decision tree generated by the classifier, Observe the confusion matrix.
   - Load each dataset into Weka and perform Naïve-bayes classification and k-Nearest Neighbour classification. Interpret the results obtained.
   - Plot RoC Curves
   - Compare classification results of ID3, J48, Naïve-Bayes and k-NN classifiers for each dataset, and deduce which classifier is performing best and poor for each dataset and justify.

5. Demonstrate performing clustering of data sets
   - Load each dataset into Weka and run simple k-means clustering algorithm with different values of k (number of desired clusters).
   - Study the clusters formed. Observe the sum of squared errors and centroids, and derive insights.
   - Explore other clustering techniques available in Weka.
   - Explore visualization features of Weka to visualize the clusters. Derive interesting insights and explain.

6. Demonstrate knowledge flow application on data sets
   - Develop a knowledge flow layout for finding strong association rules by using Apriori, FP Growth algorithms
   - Set up the knowledge flow to load an ARFF (batch mode) and perform a cross validation using J48 algorithm
   - Demonstrate plotting multiple ROC curves in the same plot window by using j48 and Random forest tree
7. Demonstrate ZeroR technique on Iris dataset (by using necessary preprocessing technique(s)) and share your observations
8. Write a java program to prepare a simulated data set with unique instances.
9. Write a Python program to generate frequent item sets / association rules using Apriori algorithm
10. Write a program to calculate chi-square value using Python. Report your observation.
11. Write a program of Naive Bayesian classification using Python programming language.
12. Implement a Java program to perform Apriori algorithm
13. Write a program to cluster your choice of data using simple k-means algorithm using JDK
14. Write a program of cluster analysis using simple k-means algorithm Python programming language.
15. Write a program to compute/display dissimilarity matrix (for your own dataset containing at least four instances with two attributes) using Python
16. Visualize the datasets using matplotlib in python.(Histogram, Box plot, Bar chart, Pie chart etc.,)