DATA MINING AND MACHINE LEARNING LAB

Course Code: CSD 421 **Credit Unit: 01**
Total Hours : 20

Course Objective:
The course will introduce data mining process which includes data selection and cleaning, machine learning techniques to ``learn'' knowledge that is ``hidden'' in data, and the reporting and visualization of the resulting knowledge. To introduce students to basic applications, concepts, and techniques of data mining. To develop skills for using recent data mining software to solve practical problems in a variety of disciplines. To gain experience doing independent study and research.

SOFTWARE REQUIREMENTS: Python, Anaconda IDE with Spider

List of experiments/demonstrations:

1. Implement and demonstrate the FIND-S algorithm for finding the most specific hypothesis based on a given set of training data samples. Read the training data from a .CSV file.
2. For a given set of training data examples stored in a .CSV file, implement and demonstrate the Candidate-Elimination algorithm to output a description of the set of all hypotheses consistent with the training examples.
3. Write a program to demonstrate the working of the decision tree based ID3 algorithm. Use an appropriate data set for building the decision tree and apply this knowledge to classify a new sample.
4. Build an Artificial Neural Network by implementing the Backpropagation algorithm and test the same using appropriate data sets.
5. Write a program to implement the naïve Bayesian classifier for a sample training data set stored as a .CSV file. Compute the accuracy of the classifier, considering few test data sets.
6. Assuming a set of documents that need to be classified, use the naïve Bayesian Classifier model to perform this task. Built-in Java classes/API can be used to write the program. Calculate the accuracy, precision, and recall for your data set.
7. Write a program to construct a Bayesian network considering medical data. Use this model to demonstrate the diagnosis of heart patients using standard Heart Disease Data Set. You can use Java/Python ML library classes/API.
8. Apply EM algorithm to cluster a set of data stored in a .CSV file. Use the same data set for clustering using k-Means algorithm. Compare the results of these two algorithms and comment on the quality of clustering. You can add Java/Python ML library classes/API in the program.
9. Write a program to implement k-Nearest Neighbour algorithm to classify the iris data set. Print both correct and wrong predictions. Java/Python ML library classes can be used for this problem.
10. Implement the non-parametric Locally Weighted Regression algorithm in order to fit data points. Select appropriate data set for your experiment and draw graphs.

Course Outcomes:

At the end of the course, the student will be able to;

- Understand the implementation procedures for the machine learning algorithms.
- Design python programs for various learning algorithms.
- Apply appropriate data sets to the machine learning algorithms.
- Identify and apply machine learning algorithms to solve real world problems.

Examination Scheme:

IA			EE			
A	**PR**	**Practical Based Test**	**Major Experiment**	**Minor Experiment**	**LR**	**Viva**
5	10	15	35	15	10	10

Note: IA –Internal Assessment, EE- External Exam, A- Attendance PR- Performance, LR – Lab Record, V – Viva.

Text Book & References:

Text Books:

- Luger George F, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 6th Edition, Addison-Wesley, 2009. (Q335.L951).
- Dunham Margaret H, Data Mining Introductory and Advanced Topics, Pearson/Prentice-Hall, 2003. QA76.9.D343D917)

REFERENCES:

- Tom M. Mitchell, Machine Learning, India Edition 2013, McGraw Hill Education.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical
- Learning, 2nd edition, Springer series in statistics.
- Ethem Alpaydın, Introduction to machine learning, second edition, MIT press.