Discipline Specific Elective (DSE) Course 4d: Advanced Statistical Computing and Data Mining

Structure 1: PG Curricular Structure with only Course Work
Structure 2: PG Curricular Structure with Course Work + Research
Structure 3: PG Curricular Structure with Research only

Course Title & Code	Credits	Credit Distribution of the Course			Eligibility Criteria	Prerequisite of the course (if any)
		Lecture (45 Hours)	Tutorial (00 Hours)	Practical (30 Hours)		
DSE 4d: Advanced Statistical Computing and Data Mining	**4**	**3**	**0**	**1**	**NIL**	**NIL**

Course Objectives:

- To introduce some advanced statistical computing techniques.
- To extract information, visualization and knowledge about various industries and finance.

Course Learning Outcomes: After successful completion of this course, student will be:

- Equipped with different theoretical methods and practicable techniques to achieve the objectives.
- Enhanced with the basic concepts of statistical theories besides developing their ability to handle real world problems with large scale data.
- Evaluate model performance using appropriate statistical and computational metrics.
- Recognize how data mining fits into the broader process of extracting useful knowledge.

Unit I (11 Hours)

Introduction to databases, tasks in building a data mining database, data warehouses, online analytical data processing, Data mining and machine learning, supervised and unsupervised learning.

Unit II (11 Hours)

Similarity and distance measures, Outliers, Minimum spanning tree, squared error clustering, K-means clustering, Hierarchical clustering, Block clustering and two-way clustering: Hartigan's block clustering algorithm, Biclustering, Plaid models for biclustering.

Unit III (11 Hours)

Extensions of regression models, McCullon-Pitts Neuron (Threshold Logic Unit), Rosenblatt's Single layer perceptron, single unit perceptron gradient descent learning algorithm, Multilayer perceptron, feed forward and back propagation learning algorithm, Self organizing maps (SOM) or Kohonen neural network, on-line and batch versions of SOM algorithm.

Unit IV (11 Hours)

Classification trees, node impurity function and entropy function, choosing the best split pruning algorithm for classification trees. Regression trees, terminal node value and splitting strategy, pruning the tree and best pruned subtree. Bagging tree based classifiers and regression tree predictors

Essential Readings:

1. Bishop, C.M. (1995). *Neural Networks for pattern Recognition*, Oxford University Press.
2. Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.

Suggested Readings:

1. Duda, R.O., Hart, P.E. and Strok, D.G. (2000). *Pattern Classification*, John Wiley & Sons.
2. Han, J. and Kamber, M. (2000). *Data Mining: Concepts and Techniques*, Morgan
3. Hand, D., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining*, MIT Press.
4. Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*, Prentice Hall. Kaufmann.
5. McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithms and Extensions*, John Wiley & Sons.
6. Nakhaeizadeh, G. and Taylor G.C., (1997). *Machine Learning and Statistics*, John Wiley & Sons.
7. Shah, N. and Shah, K. (2023). *Introduction to Data Mining*, Taylor & Francis.
8. Tufféry, S. (2024). *Data Mining and Statistics for Decision Making*, John Wiley & Sons.
9. Witten I.H., Frank, E., Hall, M.A. Hall, Pal, C.J. and Foulds, J. (2025). *Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Pub*, Morgan Kaufmann.

List of Practicals:

1. Online data processing.
2. Data mining and machine learning.
3. Clustering and unsupervised learning
4. Classification and supervised learning
5. Neural network
6. Regression trees