IE 451 – Data Visualization and Analysis
Course Syllabus

Course Instructor: Kemal Kılıç
Office: FENS 1032
Tel: 216-483 9596
E-mail: kkilic@sabanciuniv.edu

Course Description:
The course will address unsupervised learning, supervised learning, association rule mining and feature subset selection problems and introduce various techniques proposed as solutions and present their implementation particularly in the context of operations management. Data Visualization will also be introduced as part of the curriculum. Among others, probabilistic and statistical methods, clustering algorithms, classification algorithms, multiple linear regression, a priori algorithm, metaheuristics (such as genetic algorithms, simulated annealing, etc.) in the context of feature subset selection will be covered as part of the toolbox that are widely utilized in data mining.

Course Web Site: SuCourse
Lecture materials, including the class overheads, readings, assignments etc. will be available at the course web site prior to the lectures. Students are expected to check the web site regularly in order to attain the recently posted material.

Reference Texts:
TBA. Please check the SuCourse.

Marking Scheme
Assignments 20%
Midterm I 25%
Midterm II 25%
Final 30%

Assignments
There will be four assignments. Questions will be mostly implementation of the algorithms discussed in the lectures (Python).

Objection Policy
Concerns regarding marks will be accepted until two days before the last date of grades submission

Academic Conduct
Do not plagiarize other people's work. Students should be aware that anyone who engages in actions prohibited by the University’s policy on academic honesty will be subject to disciplinary action.

Course Content
1. Introduction
2. Validation
   a. Metrics for classifier evaluation
   b. Methods for classifier evaluation
c. Costs in data mining

3. Classification
   a. 0-R/1-R
   b. Naive Bayes
   c. Decision Trees
   d. Instance Based Learning (e.g. K-NN)
   e. Logistic Regression
   f. Ensemble Learning (Bagging, Randomization (e.g., Random Forest), Boosting (e.g. AdaBoost), Stacking)

4. Regression
   a. Multiple Linear Regression (MLR)
   b. Regularization in MLR (e.g., Lasso & Ridge)
   c. Classification and Regression Trees (CART) & KNN Regression

5. Clustering
   b. Hierarchical Clustering
   c. Probabilistic and Generative Clustering: Expectation – Maximization Algorithm
   d. Density Based Clustering: DBSCAN

6. Association Rule Mining
   a. Basic Terminology
   b. $A$ priori Algorithm

7. Data Understanding & Visualization
   a. Data Preprocessing
   b. Outlier Determination and Handling
   c. Imbalanced Data
   d. Missing Data
   e. Scales & Normalization of Data

8. Meta Heuristics
   a. Local Search Algorithm (i.e., Hill Climbing)
   b. Extended LSA
   c. Variable Neighborhood Search
   d. Simulated Annealing
   e. Tabu Search
   f. Genetic Algorithms
   g. Beam Search
   h. Greedy Randomized Adaptive Search Procedure (GRASP)

9. Feature Subset Selection
   a. Filtering Methods
   b. Wrappers
   c. Embedded Techniques

10. Feature Extraction (e.g. Principal Component Analysis)

11. Neural Networks
    a. Perceptron
    b. Activation Functions
    c. Feed Forward & Back Propagation
Course Schedule and Disclaimer
The course schedule can be found in ScheduleFall2023_IE451.xls which will be posted to the SuCourse. However, the instructor reserves the right, when necessary, to change examination dates, and modify the syllabus and course content. Modifications will be announced in class. Students are responsible for announced changes.