Course 9: Applied Data Analytics

Course Objective: The objective of this course is to help students develop competences on statistical techniques needed for data analysis, and various data mining techniques and algorithms used in practical problems that require processing big data for decision making purpose.

Learning Outcomes:

The students on the completion of this course would be able to

- Apply various inferential statistical analysis techniques to describe data sets and withdraw useful conclusions from the data sets (e.g., confidence interval, hypothesis testing)
- Apply data visualization techniques and key data mining techniques (e.g., classification analysis, associate rule learning, anomaly/outlier detection, clustering analysis, regression analysis) in dealing with big data sets
- Implement the analytic algorithms for practical data sets
- Perform large scale analytic projects in various industrial sectors
- Work and communicate effectively in teamwork

Prerequisite: None

Course Outline:

Module 1: Basic Data Analysis

I. Basic Concepts

1. Descriptive Statistics
2. Statistical Inferences
3. Data Measurement
4. Measures of Central Tendency and Dispersion
5. Common Statistical Graphs
6. Determination of Outliers

II. Statistical Inferences

1. Point Estimation and Required Properties of Point Estimators
2. Interval Estimations for Mean, Proportion and Variance of Population
3. Sample Size Determination

III. Hypothesis Testing

1. Hypothesis Testing for Mean, Proportion and Variance of Population – Single Sample Test
2. Hypothesis Testing for Mean, Proportion and Variance of Population – Two Samples Test
3. Type I and Type II Errors – Power of the Test
4. Observed Significance Level

Module 2: Data Visualization

IV. Data Visualization

1. Introduction to Data Visualization
2. Basic Charts for Numerical Data and Categorical Data
3. Distribution Plots
4. Multivariate Charts: Combo Chart, Combination Chart, Stacked Column Chart

V. Data Dashboard

1. What is a Data Dashboard?
2. Applications and Benefits of Data Dashboard
3. Design and Construct a Data Dashboard

Module 3: Key Data Mining Techniques

VI. Regression Analysis

1. Linear Regression and Least Square Method
2. Residual Analysis
3. Multiple Regression
4. Goodness of Fit Tests

VII. Data Classification

1. k-Nearest Neighbor Algorithm for Estimation and Prediction
2. Distance Functions: Euclidian, Manhattan, Minkowski, Min-Max Normalization, Z-Score Standardization
3. Logistics Regression
4. Bayesian Networks
5. Model Evaluation Measures for Classification Task

VIII. Data Clustering

1. Hierarchical Clustering Method
2. k-Means Clustering
3. Measuring Cluster Goodness: The Silhouette Method and The Pseudo-F Statistic

IX. Association Rules

1. Affinity Analysis
2. The a Priori Algorithm – Generating Frequent Itemsets
3. The a Priori Algorithm – Generating Association Rules
4. Measure the Usefulness of Associate Rules

X. Case Studies/Group Projects

Laboratory Sessions: None

Learning Resources:

Textbook: No designated textbook, but class notes and handouts will be provided

Reference books:

1. Larose, D.T. and Larose, C.D., Data Mining and Predictive Analytics, 2nd edition, Wiley, 2015
2. Shmueli, G., Bruce, P.C., Yahav, I., Patel, N.R. and Lichtendahl Jr., K.C., Data Mining for Business Analytics – Concepts, Techniques, and Application in R, Wiley, 2018
3. Ankam, V., Big Data Analytics, Packt, 2016
4. Walkowiak, S., Big Data Analytics with R, Packt, 2016
5. Grolemund, G., Hands-on Programming with R, O'Reilly, 2014
6. Wickham, H. and Grolemund, G., R for Data Science, O'Reilly, 2017
7. Wexler, S., Shaffer, J. and Cotgreave, A., The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios, Wiley, 2017
8. O'Cornor, E., Microsoft Power BI Dashboards Step by Step, Practice Files, 2019

Teaching and Learning Method:

The teaching is done via lectures by the instructor. Tutorial sessions are conducted on the use of tools in each subject. The learning methods include group discussion, individual/group assignment and group project/case study.

Time Distribution and Study Load:
Lectures: 30 hours
Tutorials/Group Discussions: 30 hours
Self-study: 45 hours
Group project: 40 hours

Time Allocation

Session	Activities	Time allocation
I. Basic Concepts	Lectures Other activities: · Tutorials on using Excel, R · Quizzes · Out-of-class group discussions · Individual Assignments	3 hours 6 hours
II. Statistical Inferences	Lectures Other activities: · Tutorials on using Excel, R · Quizzes · Out-of-class group discussions · Individual Assignments	3 hours 4 hours

III. Hypothesis Testing	Lectures Other activities: · Tutorials on using Excel, R · Quizzes · Out-of-class group discussions · Case Studies (in group)	4 hours 3 hours
IV. Data Visualization	Lectures Other activities: · Tutorials on using Power BI and R (ggplot2) · Out-of-class group discussions · Group Assignments	5 hours 17 hours
V. Data Dashboard	Lectures Other activities: · Tutorials on using Power BI · Out-of-class group discussions · Group Assignments	5 hours 15 hours
VI. Regression Analysis	Lectures Other activities: · Tutorials on using Minitab, R · Out-of-class group discussions · Case Studies (in group)	2 hours 4 hours
VII. Data Classification	Lectures Other activities: · Tutorials on using R · Out-of-class group discussions · Mini Group Projects	3 hours 9 hours
VIII. Data Clustering	Lectures Other activities: · Tutorials on using R · Out-of-class group discussions · Mini Group Projects	3 hours 11 hours
IX. Association Rules	Lectures Other activities: · Tutorials on using R · Out-of-class group discussions · Mini Group Projects	2 hours 6 hours

X. Group Projects	Each group is required to work on a practical dataset and analyze to withdraw insightful conclusions Activities: · Weekly group meeting · Weekly progress report · Midway report and presentation · Final report and presentation	40 hours

Evaluation Scheme: The final grade will be computed according to the following weight distribution: Mid-semester examination 20%, assignments and group projects 50%, final examination 30%. In final grading,

An “A” would be awarded if a student shows a deep understanding of the knowledge learned through home assignments, project works, and exam results.
A “B” would be awarded if a student shows an overall understanding of all topics.
A “C” would be given if a student meets below average expectation in understanding and application of basic knowledge.
A “D” would be given if a student does not meet expectations in both understanding and application of the given knowledge.

Instructor: