MIE1624HS – Introduction to Data Science and Analytics

Course description: Data science is an interdisciplinary field focused on extracting knowledge from data sets, which are typically large and applying the knowledge and actionable insights from data to solve problems in a wide range of application domains. The objective of this course is to prepare the students with the fundamental skills of data science and analytics for future study and career. In this course, we will be learning machine learning algorithms such as linear, logistic regression, decision tree, random forest, gradient boosting etc. We will also learn other skills like data mining, visualization, and text analytics. All the course projects are required to write in python and the suggested integrated development environment is Jupyter Notebook.

Course Outline

Module 1 Introduction to data science

1. Course introduction
2. Data science concept
3. Introduction to programming in python
4. Introduction to SQL

Module 2 Overview of Mathematics concept and optimization

1. Linear algebra and matrix computations
2. Derivative and convexity

Module 3 Optimization algorithm

1. Linear optimization algorithm
2. Unconstrained nonlinear optimization algorithm
3. Constrained non-linear optimization algorithm
4. Optimization case studies in Ipython

Module 4 Basic statistics

1. Random variables and Sampling

2. Probability distribution
3. Bayesian statistics
4. Hypothesis testing and statistical significance
5. Statistic case studies in Ipython

Module 5 Unsupervised machine learning

1. Hierarchical Clustering
2. K-means Clustering
3. K- Nearest neighbors
4. Principal Components Analysis

Module 6 Modeling techniques

1. Data mining
2. Preprocessing data
3. Model selection
4. Hyperparameter tuning
5. Bias-variance trade-off

Module 7 Visualization

1. Introduction to visual analytics
2. Visualization using Matplotlib
3. Visualization using Seaborn
4. Visual analytics in Tableau and Power BI
5. Visualization case studies in Ipython

Module 8 Supervised machine learning I

1. Linear regression
2. Logistic regression
3. Decision tree

Module 9 Supervised machine learning II

1. Random Forest
2. Gradient Boosting
3. K-NN
4. Naïve Bayes

Module 10 Introduction to Automated machine learning tools

1. Automated model selection
2. Automation of ML pipelines
3. Automated hyperparameter tuning

Assignments & Grading

Assignment 1 (15%): Hypothesis testing and statistical significance - individual

Assignment 2 (20%): Unsupervised machine learning Project - individual

Assignment 3 (20%): Supervised machine learning Project - individual

Group Presentation (20%)

Group Project (25%): Data analytics and business decision

All assignments should be submitted through GitHub.