€TRAINING

Data Mining and Analysis

Data Mining and Analysis

Introduction:

As the amount of research and industry data being collected daily continues to grow, intelligent software tools are increasingly needed to process and filter the data, detect new patterns and similarities within it, and extract meaningful information from it. Data mining and predictive modeling offer a means of effective classification and analysis of large, complex, multi-dimensional data, leading to the discovery of functional models, trends and patterns.

Building upon the skills learned in previous courses, this course covers advanced data mining, data analysis, and pattern recognition concepts and algorithms, as well as models and machine learning algorithms.

Course Objectives:

At the end of this course the participants will be able to:

- Analyze big data sets
- Extract patterns
- Choose the right variable impacting the results so that a new model is forecasted with predictive results.
- Design basic data collection strategies and obtain data from a number of open data sources
- Choose the right algorithms for data science problems
- Demonstrate knowledge of statistical data analysis techniques used in decision making
- Apply principles of Data Science to the analysis of large-scale problems
- Implement and use data mining software to solve real-world problems

Targeted Audience:

- Seasoned consultants wanting to transform businesses by leveraging data science and AI.
- Visionary business owners who want to harness the power of Data Science and AI to maximize revenue, reduce costs and optimize their business.
- Data Science Practitioners wanting to advance their careers and build their portfolios.
- Tech enthusiasts who are passionate about Data Science and AI and want to gain real-world practical experience.

Course Outlines:

Unit 1: Data preprocessing

- Data Cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation

Statistical inference

- Probability distributions, Random variables, Central limit theorem
- Sampling

- Confidence intervals
- Statistical Inference
- Hypothesis testing

Unit 2: Multivariate linear regression

- Specification
- Subset selection
- Estimation
- Validation
- Prediction

Classification methods

- Logistic regression
- Linear discriminant analysis
- K-nearest neighbors
- Naive Bayes
- Comparison of Classification methods

Unit 3: Neural Networks

- Fitting neural networks
- Training neural networks issues

Decision trees

- Regression trees
- Classification trees
- Trees Versus Linear Models

Unit 4: Bagging, Random Forests, Boosting

- Bagging
- Random Forests
- Boosting

Support Vector Machines and Flexible disc

- Maximal Margin classifier
- Support vector classifiers
- Support vector machines
- 2 and more classes SVMs
- Relationship to logistic regression

Unit 5: Principal Components Analysis

- ## Clustering

- K-means clustering
- K-medoids clustering

- Hierarchical clustering
- Density-based clustering

Model Assessment and Selection

- Bias, Variance and Model complexity
- In-sample prediction error
- The Bayesian approach
- Cross-validation
- Bootstrap methods