DATA SCIENCETECH INSTITUTE

MSc in Applied Data Science & Big Data

Foundations of Statistical Analysis and Machine Learning

(Machine and Deep Learning Major - DSBD2-001)

Volume of classes hours: 50 hrs *(± same personal work expected)*

Course summary:

This course introduces the students to the discipline of statistics as a science of understanding and analysing data. They will learn how to effectively make use of data in the face of uncertainty: how to collect data, how to analyse data, and how to use data to make inferences and conclusions about real world phenomena. Each of the topics is practiced through specific parts of the language, focusing on the rich set of functions and libraries available in R.

Topics to be covered are:

1. Probability and distributions
2. Foundations for inference
3. Decision errors, significance, and confidence
4. Inference for numerical variables
5. Inference for categorical variables
6. Introduction to linear regression
7. Multiple linear regression
8. ANOVA
9. Principal Component Analysis (PCA)
   & Singular Value Decomposition (SVD).

Course objectives:

- to reach sufficient proficiency so to ask - and answer to - "data" questions which are relevant to business and research.
- to allow students to orientate among the intricate “forest of data” professions and enable them to fit those within the business structure.
- to master the R language in a Data Science context to the extent of permitting independent growth of the own skills-set (in R or in competing technologies).

Course mini-projects description:

Students are required to develop a case study similar to the one proposed in class, following the data analysis pipeline.

Theoretical background used:

Students need proficiency in the topics covered during the *"Applied Mathematics for Data Science DSBD1-001"* course *(in particular, notions of calculus & linear algebra)*. However, definitions and calculation methods will be introduced as their need arise, in a convenient, self-contained and intuitive way.

Knowledge of the main office tools and previous exposure to programming will be beneficial, as per acquired during the *"Algorithmics for Data Science – Optimisation DSBD1-002"* course.

Technologies used:

A complete specification list for the machine configuration may be specified later, but at minimum, students should have Chrome or Mozilla Firefox installed: the R language, R 3.1.1 or later, RStudio. Specific libraries will be loaded and installed by the students during the course *(instructions will be provided).*