TRAINING COURSE | ONLINE

INTRODUCTION TO MACHINE LEARNING

Module I: 8-9 June 2023
Module II: 6-7 July 2023

Recent years have witnessed an unprecedented increase in the availability of information on social, economic and health-related phenomena. Today researchers, professionals and policy makers have access to enormous databases, containing detailed information on individuals, companies and institutions and use of mobile devices. Machine Learning, a relatively new approach to data analytics, which lies at the intersection between statistics, computer science and artificial intelligence, has involved rapidly over recent years in response to our need to analysis these so-called *Big Data sets*. In contrast to the more tradition approach of data analysis, which focuses on prior assumptions relating to data structure and the derivation of analytical solutions, Machine Learning techniques rely instead on a model-free philosophy development of algorithms, computational procedures, and graphical inspection of the data in order to more accurately predict outcomes. The underlying approach taken is then to "*let the data speak for itself*". Computationally infeasible until very recently, Machine Learning is itself a product of the latest advancements in both Information Technology and computing power.

This introductory course offers an intensive overview of the standard Machine Learning algorithms currently applied to social, economic and public health data, using a series of both official and community written Stata, Python and R commands. The primary objective being to illustrate how Machine Learning techniques can be applied to search for patterns in large databases, which can subsequently be used by researchers, professionals and policy makers alike to make both decisions and predictions. As a by-product, the course also serves to increase awareness as to Python and Stata's "joint" capabilities to derive knowledge and value from large and often 'noisy' databases, through the use of both official and user written Machine Learning routines developed, which to date still remain relatively unknown to the majority of users.

COURSE CODE

MODULE I: D-EF46-OL
MODULE II: D-EF47-OL

DATE AND LOCATION

The 2023 edition of this training course will be offered online. The course programme consists of 2 Modules - each divided into 2 sessions for a total of 8 hours of lessons per Module.

The first Module will be held on a part-time basis on the 8th-9th of June and the second on the 6th-7th of July both from 10 am to 2:30 pm Central European Summer Time (CEST).

COURSE STRUCTURE

The online format of our introductory Machine Learning course has been divided into two distinct modules, allowing researchers already familiar with the arguments discussed in the first module to choose to participant only in the second module.

The first module offers participants an overview of Python and Stata's Machine Learning capabilities for data management, data quality analysis; exploratory data analysis, feature engineering and Principal Component Analysis. The second module instead focuses of

the following popular Machine Learning methodologies: Supervised Learning, data management for Supervised Learning, Predictive Models, Logistic Regression, Stepwise Regression, Decision Trees, Neural Networks and Hyperparameter Optimization and Model Validation.

In common with TStat's training philosophy, each individual session is composed of both a theoretical component (in which the techniques and underlying principles behind them are explained), and an extensive applied (hands-on) segment, during which participants have the opportunity to implement the techniques using real data under the watchful eye of the course tutor. Throughout the course, theoretical sessions are reinforced by case study examples, in which the course tutor discusses and highlights potential pitfalls and the advantages of individual techniques. The intuition behind the choice and implementation of a specific technique is of the utmost importance. In this manner, the course leader is able to bridge the "often difficult" gap between abstract theoretical methodologies, and the practical issues one encounters when dealing with real data.

COURSE OUTCOMES

At the end of the course, it is expected that participants:

- have an understanding of the fundamental concepts and principles of machine learning;
- are able to use *Stata*, *R*, and *Python* for data exploration, visualization, and pre-processing data in a machine learning context;
- can independently implement the popular machine learning algorithms using *Stata*, *R*, and *Python*;
- have attained an understanding of "*real world*" data issues through this hands-on experience; and
- are able to implement solutions (with the help of the *Stata* and *Python* routine templates specifically developed for the course) to *real world* issues using Machine Learning techniques.

TARGET AUDIENCE

This course has been specifically developed for both professionals, researchers and Ph.D students working in advertising, business and management, biostatistics, economics, marketing, public health and social sciences, interested in applying the latest Machine Learning techniques in *Stata* and *Python* to big data.

PREREQUISITES

Participants are required to have a good working knowledge of:

- introduction to statistics;
- descriptive and exploratory data analysis;
- probability;
- random variables;
- probability distributions;
- sampling distributions;
- parametric estimation;

- hypothesis testing;
- simple and multiple regression.

Those needing to refresh these concepts are referred to:

- Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). A Modern Introduction to Probability and Statistics: Understanding why and how (Vol. 488). London: Springer;
- Freedman, D. (2011) Statistics. Viva Books; 4th Edition (January 1, 2011);
- McClave, James T. and Sincich, T. (2008). Statistics. Pearson;
- Ross, S. M. (2020). Introduction to probability and statistics for engineers and scientists. Academic press;
- Wasserman, L. (2004). All of statistics: a concise course in statistical inference (Vol. 26). New York: Springer.

- familiarity with a statistical software package such as SAS, Stata or SPSS is required;
- some programming experience will also be distinct advantage.

NOTE: Those wishing to participate solely in Module 2, must be confident with the arguments covered in the first Module. Due to the intensive nature of this course, the course leader will unfortunately have insufficient time to go over the material already discussed in Module 1.

PROGRAM

MODULE I | DATA MANAGEMENT FOR MACHINE LEARNING AND UNSUPERVISED LEARNING

SESSION I: DATA MANAGEMENT FOR MACHINE LEARNING AND FEATURE ENGINEERING

1. Data Management
   - Data matrix creation
   - Feature transformation (or engineering)
2. Data Quality Analysis
   - Identifying and handling missing data
   - Outliers
3. Exploratory Data Analysis
   - Subsetting data
   - Principles of Exploratory Data Analysis
4. Feature Engineering and Feature Selection
   - Feature Engineering
   - Feature Selection
5. Principal Component Analysis
   - Scree Plots
   - Biplots

SESSION II: UNSUPERVISED LEARNING

1. Unsupervised Learning
   - Introduction to unsupervised and supervised learning
2. Hierarchical Clustering
   - Choice of the approach
   - Dendrograms

3. K-Means Clustering
    - Algorithms
4. Clustering Validation
    - Clustering validation methodologies

MODULE II | SUPERVISED LEARNING

SESSION I: DATA MANAGEMENT FOR MACHINE LEARNING AND FEATURE ENGINEERING

1. Supervised Learning
    - Why supervised learning is so important
2. Data management for supervised learning
    - Feature Engineering
    - Feature Selection
3. Predictive Modelling
    - Training Set
    - Test Set
4. Linear Regression

SESSION II: UNSUPERVISED LEARNING

1. Logistic regression
2. Stepwise Regression
    - Alternatives to Stepwise Regression
3. Decision Trees
4. Neural Networks
5. Hyperparameter Optimization and Model Validation
6. Ensamble Learning

SUGGESTED READINGS

- Microeconometrics Using Stata, Volume I: Cross-Sectional and Panel Regression Methods, A. Colin Cameron and Pravin K. Trivedi, Second Edition, Stata Press (2022).
- Microeconometrics Using Stata, Volume II: Nonlinear Models and Causal Inference Methods, A. Colin Cameron and Pravin K. Trivedi, Second Edition, Stata Press (2022).
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Hastie, T., Tibshirani, R., Friedman, J., Springer (2009).
- An Introduction to Statistical Learning, Gareth, J., Witten, D., Hastie, T., Tibshirani, R., Springer (2013).
- An Introduction to R -Notes on R: A Programming Environment for Data Analysis and Graphics Version 4.2.1 (2022-06-23).
- Python 3.10.8 documentation.

REGISTRATION FEES

MODULE I: CODE D-EF46-OL (2 online sessions)
Full-Time Students*: € 475.00
Ph.D. Students: € 605.00
Academic: € 700.00
Commercial: € 940.00

MODULE II: CODE D-EF47-OL (2 online sessions)
Full-Time Students*: € 475.00
Ph.D. Students: € 605.00
Academic: € 700.00
Commercial: € 940.00

*To be eligible for student prices, participants must provide proof of their **full-time** student status for the current academic year. Our standard policy is to provide all **full-time students**, be they Undergraduates or Masters students, access to student participation rates. Part-time master and doctoral students who are also currently employed will however, be allocated academic status.

Fees are subject to VAT (applied at the current Italian rate of 22%). Under current EU fiscal regulations, VAT will not however applied to companies, Institutions or Universities providing a valid tax registration number.

The number of participants is limited to 8. Places will be allocated on a first come, first serve basis. The course will only be confirmed when at least 5 people have enrolled.

Course fees cover: teaching materials (handouts, Stata *do files* and datasets to be used during the course) and a temporary licence of Stata valid for 30 days from the beginning of the course.

Individuals interested in attending these courses must return their completed registration forms by email (training@tstat.eu) to TStat by the 29th of May to register for Module I and the 28th of June 2023 to register for Module II.

Further details regarding our registration procedures, including our commercial terms and conditions, can be found at https://www.tstattraining.eu/training/intro-machine-learning-stata-ol/.

CONTACTS

Monica Gianni
TStat Training | Kleebergstraße, 8
D-60322 Frankfurt am Main

TStat S.r.l. | Via Rettangolo, 12-14
I-67039 Sulmona (AQ)
T. +39 0864 210101

training@tstat.eu
www.tstattraining.eu - www.tstat.eu