Data Mining and Machine Learning in Spatial Analysis

Duration: 25 hours

Instructor: Joan Perez – Urban Geo Analytics

Objectives

This course aims to introduce students to the exploration of large datasets and machine learning methods, with a specific focus on spatial applications. It covers the fundamentals of data mining and both supervised and unsupervised learning, illustrated through real-world examples in urban analysis, territorial inequalities, socio-economic typologies, and urban morphology. The sessions alternate between theoretical instruction on quantitative approaches and major model families, and applied workshops using territorial data to build classifications, typologies, or predictive models.

These workshops fully integrate data visualization, thematic mapping, web scraping, and programming (in R and Python), with the goal of providing students with an operational foundation for analyzing, modeling, and representing the complexity of territories.

Course Details

The course covers a range of quantitative methods applied to geographic data, relying on the R and Python languages and specialized libraries. It begins with multivariate statistics, particularly Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA), which reduce the dimensionality of datasets while preserving their essential structure. Supervised classification is then introduced through algorithms such as k-Nearest Neighbors (KNN), decision trees, and random forests, used to predict categories from explanatory variables.

Regarding unsupervised learning, several clustering techniques are studied: k-means, hierarchical clustering, and self-organizing maps (SOM and SuperSOM), which allow the construction of typologies from multivariate data. Special attention is given to spatial clustering with the GeoClust algorithm, which introduces a geographic constraint by combining dissimilarities in variable space with spatial proximity between territorial units.

In parallel, the course places a strong emphasis on visualization of results, including exploratory graphs, interactive representations, and thematic maps. Tools such as `ggplot2`, `tmap`, `mapsf`, `leaflet`, `matplotlib`, and `geopandas` are used to coherently link the statistical

and spatial dimensions of analysis. Several clustering validation techniques are also presented, such as the silhouette index, the elbow method, and other internal indicators, to assess the relevance and stability of produced classifications.

Throughout the course, theoretical lectures alternate with practical sessions, allowing students to immediately apply the discussed concepts. Several guided exercises involve manipulating real territorial datasets to train their own models. For example, in a classification workshop, an automatic model is trained via a random forest, and students are then invited to build a manual expert system as a decision tree and attempt to outperform the machine by better interpreting variables. This approach strengthens both understanding of algorithms and the ability to use them critically.

Skills Acquired

By the end of the course, students will be able to:

- Explore and analyze large territorial databases;
- Apply multivariate techniques to identify internal structures in data;
- Implement supervised classification algorithms adapted to spatial problems;
- Build territorial typologies using unsupervised clustering methods;
- Integrate geographic constraints into analyses through dedicated methods (spatial indicators, GeoClust);
- Evaluate, compare, and validate clustering results using statistical metrics;
- Represent findings through thematic maps and analytical visualizations.

Commercial Proposal

Service title:
Training *Data Mining and Machine Learning in Spatial Analysis*
25 hours – In-person training

Instructor: Joan Perez – Urban Geo Analytics
Estimated duration: 25 hours
Total cost: Contact us

This proposal can be finalized and contracted via the **MALT** platform, which manages administrative aspects and guarantees for both parties.