

IDENTIFICATION

Course code : DAT602
ECTS : 1

DURATION

Lectures : 4
Tutorial classes : 10
Labs : 6
Total : 20

Project :
Personnal work :

ASSESSMENT

Continuous assessments and practical evaluations.

TEACHING MATERIALS

Slides, guided and practical work sheets, online learning platforms.

LANGUAGE

English

CONTACT

Kévin HOARAU
kevin.hoarau@univ-reunion.fr

Modifié le : 14 juillet 2026

UE : UE6-DATA

COURSE : Data Mining

SCOPE OF THE COURSE

OBJECTIVES

This course aims to enable students to identify the opportunities offered by data mining and explore its main application areas. It introduces the theoretical concepts necessary to understand data mining algorithms and presents practical methods for implementing these techniques on real datasets.

COMPETENCE AND SKILLS

- Exploit massive datasets using data mining techniques
- Visualize data
- Design and efficiently query a database

PROGRAM

Proximity Measurement

- Basic properties of metrics.
- Measures and special cases - Euclidean distance, Manhattan distance.
- Use of scores and rankings; desirable characteristics of scores and rankings.
- Data normalization to facilitate comparison.
- Metrics involving text
- Correlation coefficient for data sequences.
- Similarity metrics for relationships in graphs (e.g., SimRank).
- Metrics on graphs.
- Similarity of time series (e.g., DTW)

Data Preparation

- Data collection, its relation to problem-solving, the importance of specialized knowledge and openness to expert opinions
- Data sources, including databases, Internet of Things, photographs and videos, online information sources; suitability of data for objectives
- Ethical considerations regarding data acquisition and use; privacy
- Identifying potential biases in data
- Munging data - handling data errors, data gaps, cleaning data, validating data, transforming data;
- Methods for addressing dataset issues such as imbalance, insufficient attributes, and external attributes; automated and manual approaches and trade-offs between them
- The concept of features; extraction and representation of features; feature selection and feature generation

Information Extraction

- Identifying applications where information extraction plays a useful role.
- Entity and relation extraction.
- Rule-based information extraction approaches and their applications.

- Statistical-based information extraction approaches and their applications.
- Possible issues in extracted data.

Cluster Analysis

- Identifying the appropriate similarity measure for clustering activity.
- Evaluating clustering quality.
- k-means clustering algorithm, including iteration considerations.
- Density-based algorithms.
- Applications of clustering.

Classification and Regression

- Considerations for feature selection for classification
- Instance-based methods such as K-Nearest Neighbor (KNN)
- Decision tree methods
- Probabilistic models, Naïve Bayes

Pattern Mining

- The concept of association pattern mining.
- Computational complexity considerations.
- Association rule mining; Apriori and Frequent Pattern Growth (FP) algorithms.
- Sequential pattern mining; GSP algorithms.
- Efficient and parallel algorithms for pattern mining.
- Application domains

Outlier Detection

- Concept of an outlier.
- General approach - build a model of the data, then observe that a data point does not fit
- Parametric methods, such as the z-score for identifying numerical outliers in 1-D.
- Use of probability distribution functions.
- Use of depth-first approaches - after identifying the expected convex hull of a set of points, is it inside or outside; use of related graphical approaches.

Time Series

- The nature of time series data, including comparison with sequential temporal data.
- Data transformation - noise removal, normalization of time series data.
- Stationary and non-stationary time series.
- Converting time series data into discrete sequence data.
- Time series forecasting - predicting future values based on past values.
- Time series patterns - frequently encountered patterns in time series data.
- Clustering and classification of time series.
- Detecting anomalies in time series - point anomalies and shape anomalies.

BIBLIOGRAPHY

Aggarwal, C. C. (2015). Data mining : the textbook (Vol. 1). New York : Springer.

REQUIREMENTS

- Probability and random variables
- Stochastic processes
- DATA course from S5