DATA MINING 2
Course Overview

Riccardo Guidotti

UNIVERSITÀ DI PISA

Teachers

- **Riccardo Guidotti**
  - Computer Science Department
  - Email: riccardo.guidotti@unipi.it

- **Andrea Fedele (Assistant)**
  - Computer Science Department
  - Email: andrea.fedele@phd.unipi.it

Classes

- Classes
  - Monday, 09-11, Room Fib C
  - Wednesday, 11-13, Room Fib C

- Office Hours
  - Tuesday 15-17, Riccardo Guidotti's office
  - Appointment [DM2 Meeting] at riccardo.guidotti@unipi.it

- Teaching Assistant
  - Andrea Fedele [DM2 Meeting] at andrea.fedele@phd.unipi.it

No Classes and Recovery Classes

No Class

- Wed 21/02/2024
- Mon 26/02/2024
- Wed 13/03/2024
- Mon 01/04/2024 (Easter Monday)
- Mon 29/04/2024 (still not canceled)
- Wed 01/05/2024 (First of May)

Recovery Classes

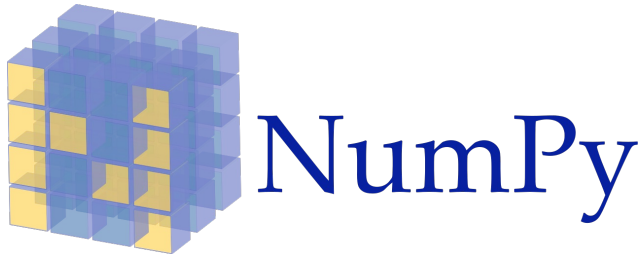
- Mon 20/05/2024
- Tue 21/05/2024
- Wed 22/05/2024
- Thu 23/05/2024 (if 29/04 is canceled)

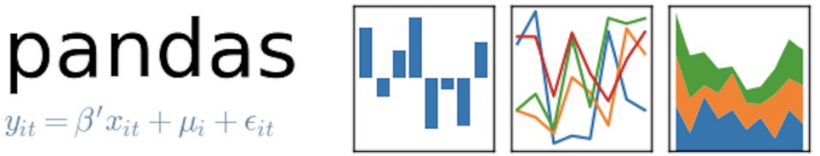
Topics

- **Module 1: Rule-based Classifiers & Transactional Data**
  - Rule-based classifiers
  - Sequential Pattern Mining
  - Transactional Clustering
- **Module 2: Time Series Analysis**
  - Time Series Similarity
  - Approximation
  - Motif, Shapelets
  - Classification, Clustering
- **Module 3: Advanced Data-Preprocessing**
  - Imbalanced Learning
  - Dimensionality Reduction
  - Anomaly Detection
- **Module 4: Advanced ML & XAI**
  - Logistic Regression
  - Support Vector Machines
  - Neural Networks
  - Ensemble Methods
  - Gradient Boosting
  - Rule-based Classifiers

Laboratory

- Python
- Jupyter Notebook

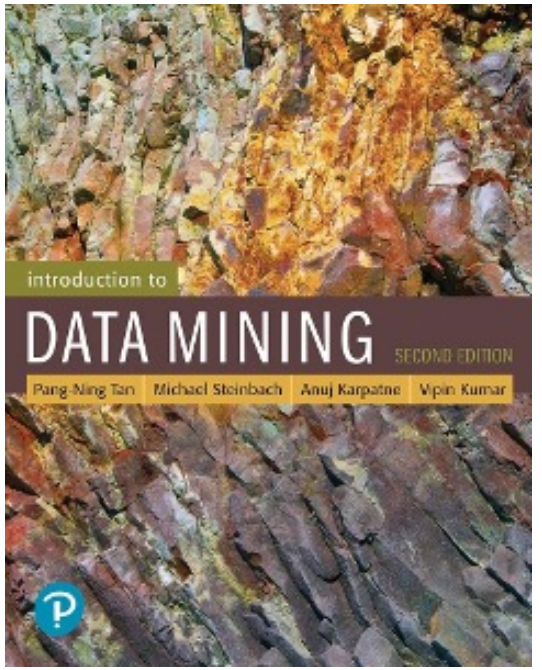




Material

- Web Site: http://didawiki.cli.di.unipi.it/doku.php/dm/start

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar. **Introduction to Data Mining**. Addison Wesley, ISBN 0-321-32136-7, 2006, 2° Edition (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)



- Berthold, M.R., Borgelt, C., Höppner, F., Klawonn, F. Guide to Intelligent Data Analysis. Springer Verlag, 1st Edition., 2010. ISBN 978-1-84882-259-7
- Laura Igual et al. Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications.
- Slides, Exercises and Notebook

Exam

- Project
  - Topics presented during the classes
  - A single report to be sent periodically and one week before the oral exam
  - Groups composed of up to 3 people (DM1), people (DM2)
- Oral Exam
  - Short discussion of the project (group presentation, where possible), plus
  - Questions on all topics presented during the classes
  - Exercises and questions about all topics

DM1 Mark = 0.6*Oral + 0.4*Project
DM2 Mark = 0.6*Oral + 0.4*Project
DM Mark = (DM1 + DM2) /2

Homework and Suggestions

Homework

- Declare Project Groups by next Tuesday 28th February adding your information at https://docs.google.com/spreadsheets/d/10R5AcqdlXsqTAxSys6zyqArvdytq4HH6Ik8Uy-NHkQ4/

- **Suggestions**
- Download and start to play with the dataset and perform data understanding.
- Use a Github repository for python and ipython files.
- Use a shared Overleaf project (LaTex) for the report.

Dataset

- **Spotify Tracks Dataset (STD) + .mp3 audio**
- The STD contains data concerning audio tracks accessible via the Spotify catalogue. These tracks span 114 distinct genres, such as hip-hop, idm, salsa, and heavy-metal. Each track is equipped with fundamental attributes: track name, artist, album name, and its popularity within the catalogue. Additionally, audio-derived features are included, encompassing aspects like danceability, energy, key, and loudness.
- The STD for the project can be found on the web page of the course.
- Detailed guidelines for the project will be presented next lecture and made on the web page of the course.

Questions?

riccardo.guidotti@unipi.it

andrea.fedele@phd.unipi.it

Let’s start!