Data mining project

Rap Italian Songs

A.Y. 2025/2026

A **project** consists of data analysis based on data mining tools. The project has to be performed by a team of 3 students. It has to be performed by using Python. The guidelines require addressing specific tasks, and results must be reported in a unique paper. This paper's total length must be **25 pages** of text including figures. The students must deliver both: paper and well-commented Python notebooks.

Download: delivered_dataset.zip

Dataset description

The data are divided in 2 csv files. The main one, *all_tracks.csv*, contains information about rap italian songs.
In the dataset there are the following variables:

1. **id**: Unique identifier for the track.
2. **id_artist**: Unique identifier for the artist.
3. **name_artist**: Name of the artist.
4. **full_title**: Full title of the song (may include artist names).
5. **title**: Title of the song only (without extra info).
6. **featured_artists**: Names of featured (guest) artists, if any.
7. **primary_artist**: The main artist of the track.
8. **language**: Main language of the lyrics.
9. **album**: Album to which the track belongs.
10. **album_name** – Name of the album.
11. **album_release_date**: Date when the album was released.
12. **album_type**: Type of album (e.g., single, album, compilation).
13. **id_album**: Unique identifier for the album.
14. **album_image**: URL/path of the album cover image.
15. **lyrics**: Full lyrics of the track.
16. **swear_IT**: Count of swear words in Italian.
17. **swear_EN**: Count of swear words in English.
18. **swear_IT_words**: List of Italian swear words found in the lyrics.
19. **swear_EN_words**: List of English swear words found in the lyrics.
20. **n_sentences**: Number of sentences in the lyrics.
21. **n_tokens**: Number of tokens (words) in the lyrics.
22. **tokens_per_sent**: Average number of words per sentence.
23. **char_per_tok**: Average number of characters per token (word length).
24. **lexical_density**: Ratio of unique words to total words (measure of vocabulary richness).

25. **avg_token_per_clause**: Average number of tokens per clause.
26. **year**: Release year of the track.
27. **month** – Release month of the track
28. **day** – Release day of the track.
29. **bpm**: Tempo of the song, measuring how many beats occur in one minute. Higher values = faster tempo.
30. **centroid**: Spectral centroid, i.e., the "center of mass" of the spectrum. It's often perceived as the **brightness** of the sound (higher centroid → brighter/tinnier timbre; lower → darker/warmer).
31. **rolloff**: Spectral rollof, i.e., the frequency below which a certain percentage (often 85–95%) of the total spectral energy lies. Used to separate harmonic content from noise.
32. **flux**: Spectral flux measures the rate of change in the power spectrum between consecutive frames. High flux means the sound is changing rapidly (e.g., percussive or noisy).
33. **rms**: Root Mean Square energy, i.e., overall energy or loudness of the audio signal. Higher RMS → stronger intensity.
34. **zcr**: Zero Crossing Rate, i.e., the rate at which the signal changes sign (crosses zero). Higher ZCR → noisier or more percussive sounds; lower → smoother/tonal sounds.
35. **flatness**: Spectral flatness describes how "flat" the spectrum is. High flatness → noise-like signal (all frequencies equally present), low flatness → tonal/harmonic signal.
36. **spectral_complexity**: Counts how many peaks are present in the spectrum. Higher complexity means the sound has more simultaneous harmonic or tonal components.
37. **pitch**: The fundamental frequency of the audio (perceived as the musical note).
38. **loudness**: Perceived volume of the track (not just amplitude, but weighted to human hearing perception).
39. **disc_number**: The disc number (for multi-disc albums).
40. **track_number**: The track's number within the album.
41. **duration_ms**: Duration of the track in milliseconds.
42. **explicit**: Whether the track is marked as explicit.
43. **popularity**: Popularity score of the track (usually 0–100).
44. **stats_pageviews**: Number of visualizations of the song.
45. **modified_popularity**: Data may come from different sources: the popularity score has been adjusted accordingly, this flag indicates the presence of such adjustments.

The second file, *artist.csv* contains information about the artists of the songs included in tracks.csv. It includes the following variables:

1. **id_author**: Unique identifier for the artist/author.
2. **name**: Full name of the artist.
3. **gender**: Gender of the artist.
4. **birth_date**: Date of birth of the artist.
5. **birth_place**: Place of birth (usually city or town).
6. **nationality**: Nationality of the artist.
7. **description**: Short biography or description of the artist.
8. **active_start**: Date when the artist started their professional activity.

9. **active_end**: Date when the artist ended their professional activity (or missing if still active).
10. **province**: Italian province where the artist is associated.
11. **region**: Italian region corresponding to the artist's province.
12. **country**: Country associated with the artist of artist's origin.
13. **latitude**: Geographic latitude of the location of artist's origin.
14. **longitude**: Geographic longitude of the location of artist's origin

Task1: Data Understanding and Preparation (30 points)

Task 1.1: Data Understanding

Explore the Rap Italian Song dataset with the analytical tools studied and write a concise "data understanding" report assessing data quality, the distribution of the variables and the pairwise correlations.

Task 1.2: Data Preparation

Improve the quality of your data and prepare it by extracting new features interesting for describing the songs. Therefore, you are going to describe the single incident and examples of indicators to be computed are:

- Relative popularity with respect to the artist, album, or year.
- Rank songs by pageviews per year
- Relative audio features of a song with respect the songs of an author
- Popularity of a song with respect to the ones of the same region
- For each artist, compute the entropy of the distribution of the audio's features across their songs, so you can measure variation in musical/lyrical characteristics

Note that these examples are not mandatory. You can derive indicators that you prefer and that you consider interesting for describing the incidents.

It is MANDATORY that each team defines some indicators. Each of them has to be correlated with a description and when it is necessary also its mathematical formulation. The extracted variables will be useful for the clustering analysis (i.e., the second project's task). Once the set of indicators is computed, the team has to explore the new features for a statistical analysis (distributions, outliers, visualizations, correlations).
Subtasks of DU:

- Data semantics for each feature that is not described above and the new one defined by the team
- Distribution of the variables and statistics
- Assessing data quality (missing values, outliers, duplicated records, errors)
- Variables transformations
- Pairwise correlations and eventual elimination of redundant variables.

Nice visualization and insights can be obtained, exploiting the latitude and longitude features (e.g. https://plotly.com/python/getting-started/).

Task 2: Clustering analysis (30 POINTS - 32 with optional subtask)

Based on the features extracted in the previous task, explore the **song** dataset using various clustering techniques. Carefully describe your decisions for each algorithm and which are the advantages provided by the different approaches.

Subtasks

- Clustering Analysis by K-means on the entire dataset:
    1. Identification of the best value of k
    2. Characterization of the obtained clusters by using both analysis of the k centroids and comparison of the distribution of variables within the clusters and that in the whole dataset
    3. Evaluation of the clustering results
- Analysis by density-based clustering on the entire dataset:
    1. Study of the clustering parameters
    2. Characterization and interpretation of the obtained clusters
- Analysis by hierarchical clustering on the entire dataset:
    1. Compare different clustering results got by using different version of the algorithm
    2. Show and discuss different dendrograms using different algorithms
- Final evaluation of the best clustering approach and comparison of the clustering obtained
- **Optional (2 points):** Explore the opportunity to use alternative clustering techniques in the library: https://github.com/annoviko/pyclustering/

Task 3: Predictive Analysis & XAI (32 POINTS)

Consider the problem of predicting, for each song, the **school of origin of the singer**.

- In this context, a school groups artists who share similar characteristics (musical style, lyrical content, production patterns, etc.). Artists belonging to the same school are expected to have comparable traits.
- To define these schools, you can use the region of the artist to create meaningful macro-zones, which will act as class labels.

The students need to:

1) design your own mapping from regions to macro-zones (schools) for **defining the class labels**, but the task must be formulated as a multi-class classification problem. An example of mapping may be:
    - North: Lombardia, Piemonte, Veneto, Liguria
    - Campania
    - Lazio
    - Sardegna
    - Center: Toscana, Emilia-Romagna, Marche
    - South: Puglia, Sicilia, Calabria

2) define new features that enable the classification. Please, reason on the suitability of the features defined for the clustering analysis. In case features, used for clustering analysis, are not suitable for the above prediction problem you can also change them.
3) perform the predictive analysis comparing the performance of different models, discussing the results and the possible preprocessing applied to the data for managing potential identified problems that can make the prediction task hard. Note that the evaluation should be performed on both training and test sets.
4) Explore the opportunity to apply explainable AI methods for adding a transparency layer for complex predictive models. You can use the following libraries discussed during the lectures:
    - https://github.com/interpretml/interpret for EBM
    - https://github.com/slundberg/shap for SHAP
    - https://github.com/marcotcr/lime for LIME

Task 4: Time Series Analysis (32 POINTS)

Consider the song dataset and only songs by "**Fedez**" and "**Fabri Fibra**".
Explore the mp3 files exploiting *Librosa* Python package, designed for music and audio analysis. You can exploit it together with *tslearn* to extract features. Librosa is designed more for specific audio analysis, while tslearn for general purpose time series analysis.
https://librosa.org/doc/latest/index.html

Task 4.1: Clustering and motif/anomalies extraction

The goal of this task is grouping similar songs through the use of the mp3 audios. Analyze the results of the clustering and extract motifs and anomalies in the time series for a deep understanding and exploration.

Task 4.2: Shapelet extraction

Exploiting the mp3 time series, extract the shapelet according to the song authorship attribution, i.e., *Fedez* and *Fabri Fibra*.

In the following URL, you will find the MP3 files of Fedez and Fabri Fibra's songs.
Download: MP3 dataset

Task 5: Ethical and legal implications (30 POINTS)

Students need to analyse the data analytics workflows and discuss the potential ethical and legal implications derived from the access and use of data, from the deployment and application of the learned models. Moreover, in case some ethical and legal risks are identified, discuss potential strategies that could be defined for addressing the identified issues.

This analysis should cover at most 2 pages in the report. Please check the updated version of the next section.

Rules for final delivery and Exam

Project Delivery. The final deadline of the project is **5th January 2026 at 23:59**. This deadline is **STRICT**. No extension is possible because then the winter session of exams starts. Groups that will not deliver the project by 5th January will need to do the written exam. Each group must deliver by email to anna.monreale@unipi.it, mattia.setzu@unipi.it, lorenzo.mannocci@di.unipi.it a zipped folder named **DM_GroupID.zip** and containing 4 folders and 1 pdf file:

1. a folder named **DM_GroupID_TASK1**, containing source code of data understanding
2. a folder named **DM_GroupID_TASK2**, containing source code of data clustering
3. a folder named **DM_GroupID_TASK3**, containing source code of classification and explanation analysis
4. a folder named **DM_GroupID_TASK4**, containing source code of time series analysis
5. a pdf file with maximum 25+2 pages (25 pages for tasks 1-4 and 2 pages for task5) including figures discussing the results of the tasks. The name of this file must be: **DM_Report_GroupID.pdf.** The file must contain the list of authors (i.e., members of the group).

The **subject** of the email must be **"DMProject25_GroupID"**

How to book for the exam colloquium?
In https://esami.unipi.it/ you can find the dates for the exam: one for January and one for February. Each student must do the registration on one of the 2 dates. These are not the dates of the colloquium but we will use the list of registered students for organizing the exam dates. We will share with you a calendar for the oral exam.