Challenge 2: supervised learning

Music Classification and Regression

Context

The 2025 Challenge 2 is inspired by real-world music information retrieval problems and uses a dataset built from **track-level metadata** and **audio descriptors.**

Music platforms rely heavily on machine learning models to organize catalogs, classify tracks into genres, and recommend content. In this challenge, you will use a large, heterogeneous dataset to build predictive models for musical genres and track duration.

Your objective is to explore the dataset thoroughly, preprocess it properly, and implement the supervised learning methodology seen in class by addressing three predictive tasks.

Data description

You are provided with **four datasets**. The first three datasets describe music tracks. The fourth one is about musical genres.

1. `tracks.tsv`

Track metadata including:

- `track_id`
- artist information (name, ID, location, latitude/longitude)
- album information
- listening statistics: `interest`, `listens`
- primary genre label: `genre_top`
- duration of the track (in seconds)

2. `echonest_features.tsv`

High-level audio descriptors typically used in music classification such as danceability, energy, tempo etc.

3. `spectral_features.tsv`

Low-level audio features focusing on spectral statistics (frequency distribution), such as skew, bandwidth, kurtosis, mean, stdev, etc.

4. `genres.csv`

An additional hierarchical genre taxonomy that you may use to improve your models.

Work to be done

Phase 1: warm-up (2 points)

The goal of this phase is to understand, clean, and prepare the data before building prediction models.

Descriptive statistics and feature engineering

1. Merging the datasets, cleaning the columns names
2. Data quality assessment (missing values, duplicates, inconsistencies)
3. Uni- and multi-dimensional statistics: understanding of structure, links between variables, etc.
4. Variable recoding, transformation, possible creation of new variables

Phase 2: Construction of prediction models (6 points)

You will build **supervised learning models** for three tasks:

Task 1 — Predict the original genre (`genre_top`) – 2 points

A multi-class classification problem. Try to reach the best performance level and also explain possible issues.

Task 2 — Predict your coarse-grained genre (3–4 categories) – 2 points

You must:

- define your own taxonomy of coarse genres (3 or 4 categories),
- justify how and why genres were grouped,
- train several models and compare performance with Task 1.

This task aims to examine whether fewer classes lead to better generalization.

Task 3 — Predict the track duration – 2 points

Try to build the best regression model to predict the track duration. Design appropriate feature sets (metadata only? audio only? both?) and evaluation metrics.

For each of the three tasks, you must test a broad set of models:

  - Logistic Regression
  - kNN
  - SVM
  - Decision Trees & Random Forests
  - Boosting
  - Gradient Boosting / XGBoost / LightGBM
  - Neural Networks / Deep Learning (optional)

Evaluation criteria:

Each team is fully responsible for:

- designing their train/validation/test splits,
- designing rigorous evaluation pipelines
- preventing data leakage
- justifying their validation strategy
- demonstrating good ML practices and reproducibility

Phase 3: Code cleaning and report writing

The quality of your code and your report (well commented notebook and/or PDF report) will be considered (2 points).

The use of generative AI (LLM) for writing your report is prohibited. In case of doubt, we reserve the right to making each team member undergo a final oral exam.

DEADLINE: DECEMBER 5th, 11.30pm