# Machine Learning Basics

*Instructors: Babak n. Arabi, Mohammadreza A. Dehaqani, Mostafa Tavassolipour*

Mostafa Kermani Nia, Faezeh Mozaffari, Mahan Osouli

Fall 2025

## Final Project

## Introduction

Imagine listening to a short audio clip and being able to recognize the language instantly, even without understanding a single word. Behind this seemingly simple ability lies a rich structure of acoustic patterns shaped by phonetics, rhythm, and intonation. In Machine Learning, uncovering such hidden patterns is a central challenge especially in the domain of speech and audio processing.

In this project, you step into the role of a Machine Learning practitioner tasked with teaching a computer to distinguish between spoken languages. Your journey begins with collecting raw audio data from selected audiobook sources in multiple languages. By exploring these recordings, you will discover how subtle differences in sound can be captured through carefully designed audio features. Using these features, you will then build and analyze models that can automatically identify the language spoken in short audio segments.

This project guides you through the complete Machine Learning pipeline, from data collection and feature extraction to the application of classification and clustering techniques. Along the way, you will strengthen your ability to think critically about data, make informed modeling choices, and address real-world challenges in speech-based Machine Learning systems.

## Project Description

In recent years, the analysis and processing of speech signals have played a critical role in many real-world applications such as spoken language identification, speaker recognition, emotion recognition, and speech-based intelligent systems. Speech signals inherently contain rich and meaningful information in both time and frequency domains, which can be leveraged for machine learning–based classification tasks.

In this project, the primary focus is on Spoken Language Identification (SLID) and the investigation of language-dependent acoustic characteristics. Unlike conventional projects where a clean, preprocessed dataset is provided, students are required to actively participate in the dataset creation process.

In this phase, students must collect speech samples from specified audiobook sources. It is essential that the selected audio segments inherently exhibit the characteristics relevant to language discrimination. The goal is not merely to apply preprocessing techniques afterward, but rather to carefully select voice recordings that already possess the desired linguistic and acoustic properties.

Students are strongly encouraged to listen carefully to all collected audio samples and ensure that:

- The speech is clear and representative of the target language
- The recording quality is sufficient for extracting meaningful language-related features
- The audio content reflects natural linguistic patterns

This step is crucial, as the success of later stages depends heavily on the quality and suitability of the collected speech signals, not just on signal preprocessing.

## Dataset

In this phase, the entire class will collaboratively build a multilingual speech dataset that will be used in subsequent phases of the project. Each student is responsible for collecting and documenting a set of speech

recordings and providing detailed metadata for each audio file.

1. **Sign-up Sheet**

   To maintain a balanced distribution of languages, we have prepared a Shared Sign-up Spreadsheet containing specific audiobooks and podcasts. You must access this sheet, select an available row that specifies the target Language and Source Name, and write your Name and Student ID in that row to claim it. Additionally, make sure to enter the start and end times for the section you choose from the selected source in the corresponding columns of the sheet to prevent any overlap of data. You are strictly required to collect data only from the source you have assigned to yourself. This spreadsheet is used to document essential information about the collected speech samples and must be filled out before submission.

2. **Audio Collection**

   You must extract 15 to 20 audio files in `.mp3` format. Each file should be approximately 1 minute long. It is critical that you respect sentence boundaries. Every audio file must begin exactly at the start of a sentence and end exactly after a sentence finishes. Do not cut the audio in the middle of a phrase or word. The audio should contain clear speech from the narrator; avoid segments with overpowering background music or long periods of silence.

3. **File Naming**

   To organize your data, you must rename all audio files using your Student ID in the format `SID_male/female_targetlang_voice#.mp3`, then complete the Shared Multilingual Audio Dataset Spreadsheet. There are different sheets available in this link, each corresponding to a different source. Students must ensure that the information entered in the spreadsheet accurately corresponds to the submitted audio files. Incomplete or inconsistent entries may result in the exclusion of the data.

4. **Submission**

   Place all your renamed `.mp3` files into a single folder, compress it into a ZIP file, and upload it on the eLearn platform. Do not include any other files. Ensure your filenames are correct. If the Student ID in the filename does not match the ID in the Shared Multilingual Audio Dataset Spreadsheet, your data cannot be identified and will be discarded.

## Report

The report must include the following sections:

1. **Introduction to Language Identification**

   - Definition of Spoken Language Identification and its importance in applications such as voice assistants and multilingual systems
   - Comparison between speech data recorded in controlled environments and real-world data such as audiobooks
   - Explanation of Closed-set LID and Open-set LID, along with their conceptual and practical differences
   - A brief overview of implementation approaches for each paradigm

2. **Challenges in Language Identification**

   Identification and explanation of key challenges in LID, including:

   - Phonetic similarities between related languages
   - Accent and speaker variability
   - Short-duration speech segments

   Review of existing research directions and solution strategies addressing these challenges

3. **Audio Data Characteristics and Selection**
   - Discussion of why proper audio selection is critical for language identification
   - Explanation of desirable properties of speech samples for LID tasks
   - The role of listening-based qualitative evaluation alongside technical analysis

4. **Feature Extraction Techniques**

   Explanation of various feature extraction methods in speech-based language identification, including but not limited to:
   - Mel-Frequency Cepstral Coefficients (MFCC)
   - Fast Fourier Transform (FFT)
   - Log-Mel Spectrogram
   - Spectral Centroid
   - Chroma Features
   - Spectral Contrast
   - Zero-Crossing Rate
   - Linear Predictive Coding (LPC)
   - Perceptual Linear Prediction (PLP)

   For each method, students must:
   - Explain how the method works
   - Analyze whether it is suitable for distinguishing languages
   - Discuss its advantages and limitations with references to credible sources

5. **Similarity Learning**
   - Definition of Similarity Learning in speech and language analysis
   - Introduction to common loss functions such as Contrastive Loss and Triplet Loss
   - Discussion of how similarity learning can be applied to language discrimination

## Group Formation

For Phase 2 of the project, students are required to register their groups using the Group Registration Spreadsheet. Each group should consist of two or three members. Only one representative from each group must complete the spreadsheet on behalf of all group members by entering the names of their teammates and degree level (Bachelor or Master).

## Final Notes

Beyond strictly following the project instructions, students are encouraged to think and act as Machine Learning practitioners. If certain aspects of the project are not fully specified, students should independently research relevant methods, justify their choices, and make informed decisions. While guidance and support will be available throughout the course, initiative, critical thinking, and methodological reasoning are highly valued. Consider this project as an opportunity to go beyond grades and focus on meaningful learning, effort and thoughtful work will always be recognized.