AICE 2002 – Assignment #1

- ★ This assignment counts for 20% of your module mark
- ★ Total possible: 100 points
- ★ Due date: **November 6, 2025 at 12:00pm** (10% late penalty per day)
- ★ Submit your assignment electronically on Moodle in **one single zip file**
- ★ Submit: code (as Python .py scripts or as a Jupyter Notebook) and a short report as a PDF (5-6 pages maximum), clearly marked where each task has been addressed.

Dataset: You are given a manageable portion of a larger dataset consisting of features extracted from audio files. The audio files come from 15 different male speakers, and 15 different female speakers. The audio files originate from three different sources ***System_Orig, System_X,*** and ***System_Y***. System_orig is original audio as it was first recorded. System_X and System_Y are two different AI systems where each AI system has made different modifications to the audio prior to feature extraction. Each entry in the dataset corresponds features from a single audio file, and includes a set of labels:

- ***sid*** – This label is the anonymised ID of the person who spoke in the audio. All speaker labels begin with the letter *p* followed by three digits (e.g., *p225*). These labels came with the original speech dataset when it was first collected.
- ***uid*** – This label is the utterance ID that links the data point to a specific sentence that was spoken. Each speaker contributes a set of utterances, some of which may overlap across speakers.
- ***gender*** – This label is a binary categorisation of male/female speakers, and these labels came with the original speech dataset when it was first collected.
- ***target_human*** – This label is a nominal categorical label with four possible options (e.g., *lamb, goat, sheep, wolf*). These labels were added to the dataset using human perception surveys.
- ***target_asv*** – This label is similar to the *target_human* labels. The main difference is that the labels were obtained through an AI analysis technique called automatic speaker verification (ASV) scoring, rather than from human perception surveys.

Assignment: Which labelling scheme (target_human or target_asv) is more challenging to predict from the given data? Design and execute an experiment that compares classification performance for predicting target_human vs. target_asv labels. You must compare performance with at least two different algorithms in your solution. You may design your experiment to use only supervised ML, an ensemble, or a combination of supervised/unsupervised ML at your discretion, to obtain the best possible prediction performance. Report classification results in a table with at least two different performance metrics. You should include additional tables, figures, and plots to support and explain your strategic experimental decisions and outcomes.

In your report, for Tasks 1-6 indicate which lines of code correspond to each task and also describe:

Task 1) How you divided your dataset into train/valid/test splits and why. How are the label distributions represented in each split? Did you combine data from all three systems to make a larger dataset or keep them separate? Did you divide the data by gender or speaker? What steps did you take to make the problem more tractable? Did you use k-fold cross-validation, and if so how many folds?

Task 2) How you pre-processed the dataset and why. Did you normalise or standardise any of the features? Did you have missing values and if so, how did you handle these? Are the classes balanced or imbalanced and did you take any steps to alter the class balance? Did you add noise to the dataset, and if so, why and what kind of noise?

Task 3) Which algorithms you chose and why. Did you use a mix of supervised and unsupervised? Did you perform any feature selection or visualisation before selecting an algorithm?

Task 4) How you handled the data across the three systems (orig, X, and Y). Are you reporting results for each system independently or combined as one large dataset? What are the benefits or tradeoffs from your decision?

Task 5) Which evaluation metrics you chose to inform your conclusions and why. Which evaluation metrics were most useful for answering the research question? Is performance on one evaluation metric better than performance on another? Is performance on a particular category label better than another (e.g. easier to predict *lamb* and more challenging to predict *wolf*)?

Task 6) What range of hyperparameters you tried for optimal performance from your selected algorithms. Some algorithms have parameters that can be 'swept' by running additional experiments to find the optimal combination of tunable parameters. If you tried different parameters explain this, report the values in a table, and indicate which ones performed best.

Task 7) A few sentences to summarise your findings and conclusions. Explain how your *experiment design* and your *experiment results* answer the original research question. Is the assignment question answerable/solvable, why or why not? Does one algorithm perform better than another, why or why not? Are there other experiments that you would run, or other ways you would test your design, if you had more time?

Marking:

- Correctness of code, readability, plots = 30 points
- Clarity and quality of report = 10 points for each task (total of 70 points possible)
    - Clearly mark where in the report each task is addressed (e.g. use subheadings or other indicators to make this clear for marking purposes)
    - Please use IEEE conference style in Overleaf (template will be provided) but submit the report as a compiled PDF file - NOT as a LaTeX file.
    - Maximum 5-6 pages (not including references, which are optional)