INF8245E – Machine Learning (Fall 2025)
Kaggle Competition Instructions

1 Problem Definition

Antimicrobial resistance (AMR) is one of the greatest challenges in modern medicine. Through mutations of their genome, individuals of a bacterial species can gain mechanisms that make antibiotics ineffective, threatening our ability to treat even common infections. Understanding and predicting this resistance from genomic information is a crucial step toward developing better diagnostics and treatment strategies.

In particular, Escherichia coli is one of the most widely studied bacterial species worldwide thanks to its ease of manipulation. In spite of this, we are still not able to fully understand the mechanisms that make it resistant or not to certain antibiotics.

In this competition, your goal is to predict whether an Escherichia coli individual is resistant or susceptible to a given antibiotic, based on its genomic sequence. The dataset consists of a training set (with labels) and a test set (without labels). You must submit your predictions for the test set on Kaggle. The performance metric is the macro-averaged **F1-Score**[1].

2 Team Formation

First, you need to create an account on the Kaggle website, if you haven't already. Next, you can access the competition. We expect you to be working in groups of exactly 3. To be able to form a team, follow the instructions below:

- Each team should consist of exactly 3 members.
- Fill out this Google form (https://forms.gle/K85MVXConk2rEMTv6) with your team's information by October 28th at 11 PM EST.
- Register as an individual Kaggle user, enter the competition, and accept the terms and conditions.
- Go to the Kaggle team section. In *Invite Others*, enter your teammates' names, or team name, and request a merge.
- Your teammate has the option to accept your merge. The person accepting a merger is the team leader.

Note on number of submissions: The maximum number of submissions is 2 per day for the entire team. The data will be released after the team formation deadline. All the team members will receive the same marks for this competition. It is your responsibility to ensure that everyone has contributed equally to the competition.

3 Instructions

To participate in the competition, you must provide a list of predicted outputs for the instances on the Kaggle website. To solve the problem, you are encouraged to use any classification

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

methods you can think of, presented in the course or otherwise. Note: We suggest that you start early, allowing yourself enough time to submit multiple times and get a sense of how well you are doing.

4 Report

In addition to your methods, you must write up a report that details the pre-processing, validation, algorithmic, and optimization techniques. It should also report your exact Kaggle results. The report should contain the following sections and elements:

- **Project title**
- **Team name** on Kaggle, as well as the list of team members, including full names, email addresses, and matricules.
- **Feature design:** Describe and justify your pre-processing methods, and how you designed and selected your features.
- **Algorithms:** Give an overview of the learning algorithms used without going into too much detail, unless you judge it necessary.
- **Methodology:** Include any decisions about training/validation split, distribution choice for Naive Bayes, regularization strategy, any optimization tricks, setting hyper-parameters, etc.
- **Results:** Present a detailed analysis of your results, including graphs and tables as appropriate. This analysis should be broader than just the Kaggle result: include a short comparison of the most important hyperparameters and all the methods you implemented.
- **Discussion:** Discuss the pros/cons of your approach methodology and suggest areas of future work.
- **Statement of Contributions:** Briefly describe the contributions of each team member towards each of the components of the project (e.g., defining the problem, developing the methodology, coding the solution, performing the data analysis, writing the report, etc.). At the end of the Statement of Contributions, add the following statement: "We hereby state that all the work presented in this report is that of the authors."
- **References** (optional)
- **Appendix** (optional): Here you can include additional results, more details of the methods, etc.

The main text of the report should not exceed 6 pages. References and appendix have no page limit. You should use the ICLR format, whose template you can find online[2].

5 Submission Requirements

We expect you to follow these rules:

- You must submit the code developed during the project. The code must be well-documented and include a README file containing instructions on how to run it. A form to submit the code will be released at the end of the competition.

[2] https://github.com/ICLR/Master-Template/raw/master/iclr2026.zip

- Your submission folder should contain a notebook named ”final.ipynb”, which reproduces your predictions exactly. Make sure to fix the random seeds so that the generated predictions exactly match your submitted prediction file. Also, make sure we can directly run it without making any modifications.
- The prediction file must be submitted online at the Kaggle website. Please make sure your submitted result file has the correct structure and format. You should submit your result in .csv format. More information about the correct structure and format can be found on the Kaggle website (go to: Overview → Evaluation).
- You must submit a written report according to the general layout described above. The report must be submitted in GradeScope under "Kaggle Report". The competition ends on December 5th, and the report is expected by December 8th. The late submission policy is the same as the default policy used for the other assignments.

6 Evaluation Criteria

Marks will be calculated with 40% for performance on the private test set in the competition and 60% for the written report. For the competition, the instructors will set up a baseline and compute the grades as follows:

- The highest scoring team will get 100%.
- All other teams' grades will be determined by scaling their score relative to the highest score.

For the written report, the evaluation criteria include:

- Technical soundness of the methodology (pre-processing, feature selection, validation, algorithms, optimization).
- Technical correctness of the description of the algorithms (may be validated with the submitted code).
- Meaningful analysis of final and intermediate results.
- Originality of the approach.
- Correct insights and analysis on the link between certain attributes and the target.
- Clarity of descriptions, plots, figures, and tables.
- Organization and writing. Please use a spell-checker and don't underestimate the power of a well-written report.

Do note that the grading of the report will emphasize the rationale behind the pre-processing and optimization techniques. The code should be clear enough to reflect the logic articulated in the report. We are looking for a combination of insight and clarity when grading the reports.

7 Questions and Clarifications

For additional questions, please use **Piazza** or attend **TA office hours**.