INF8245E(Fall 2024) : Machine Learning Given on: Nov 7^{th}

Kaggle Competition Due on : December 8th, 10 pm

- The Kaggle competition will be done in groups. We expect you to be working in groups of exactly 3.
- You must submit the code developed during the project. The code must be well documented. The code should include a README file containing instructions on how to run the code. Submit the code as a zip file in GradeScope under ”Kaggle Code”.
- When Submitting to GradeScope, be sure to submit
  1. A ‘.zip‘ file containing all your python codes as well as a final notebook named as ”final.ipynb” to the ‘Kaggle Code‘ section on Gradescope.
  2. A ‘pdf‘ file of your report that details the pre-processing, validation, etc. to the ‘Kaggle Report‘ section on Gradescope.
- More details on the instructions, submissions and report will be detailed below.
- If you have questions regarding the assignment, you can ask for clarifications in Piazza. You should use the corresponding tag for this assignment.
- Before starting the assignment, make sure that you have registered and joined a team on the Kaggle Competition page. Also make sure to have downloaded all the data for the competition and put them in the appropriate locations.
- Throughout the assignment, you will need to explore different techniques and algorithms to improve the performance on the task. Make sure you document and provide clear code for reproducibility.
- You cannot use ChatGPT or any other code assistance tool for the programming part, however if you use ChatGPT to edit grammar in your report, you have to explicitly state it in the report.

Problem Statement

In this Kaggle competition, you will be tasked with a Diabetes diagnosis task. It is known that a person’s lifestyle have an influence on a patient’s health. The goal of the task is then to better understand the relationship between lifestyle and diabetes. The dataset provided contains 200K records and 28 features that consists of demographics, lab test results, and answers to survey questions. The target variable for classification is whether a patient has diabetes or not. The files for the train and test set will be available on the Kaggle competition website.

For the competition, we will provide a training set (with labels) and a test set (without labels). Your predictions on the test set will be submitted to Kaggle and we will use F1-Score as the performance metric.

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:

1. Each team should consist of exactly 3 members.
2. Fill out this Google form (https://forms.gle/gTREM4YNH4DTcGQv6) with your team's information by Nov 6th at 10 PM, EST.
3. Register as an individual Kaggle user, enter the competition and accept the terms and conditions.
4. Go to the Kaggle team here: https://www.kaggle.com/t/e7f9d1a8ba594ddb83d7ef7f6bcc4e8d
5. In the Invite Others section, enter your teammates' names, or team name.
6. 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 amount of submissions is 2 per day for the entire team. The test data will be released after the team formation deadline which is November 7th at 10 PM EST. If you make any submissions before this, you might be disqualified from the competition.

All the team members will receive the same marks for this competition. It is your duty to make sure everyone has contributed to the competition equally.

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 methods you can think off, presented in the course or otherwise. Looking into creative way to create new features from those provided may prove especially usefull in this competition. **Note**: We suggest you to start early, allowing yourself enough time to submit multiple times and get a sense of how well you are doing.

Report

In addition to your methods, you must write up a report that details the pre-processing, validation, algorithmic, and optimization techniques, as well as providing your Kaggle results that we compare them with. The report should contain the following sections and elements:

1. Project title
2. Team name on Kaggle, as well as the list of team members, including full names, email and matricules.
3. Feature design: Describe and justify your pre-processing methods, and how you designed and selected your features.
4. Algorithms: Give an overview of the learning algorithms used without going into too much detail in the class notes (e.g. SVM derivation, etc.), unless you judged necessary.
5. Methodology: Include any decisions about training/validation split, distribution choice for Naive Bayes, regularization strategy, any optimization tricks, setting hyper-parameters, etc.
6. 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 hyper- parameters and all the methods you implemented.
7. Discussion: Discuss the pros/cons of your approach methodology and suggest areas of future work.
8. 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.”
9. References (optional).
10. Appendix (optional). Here you can include additional results, more detail of the methods, etc.

The main text of the report should not exceed 6 pages. References and appendix can be in excess of the 6 pages. You should use the ICLR format `https://www.overleaf.com/latex/templates/template-for-iclr-2025-conference-submission/gqzkdyycxtvt`. You can find the template online.

Submission Requirements

We are expecting you to follow these rules:

1. You must submit the code developed during the project. The code must be well documented. The code should include a README file containing instructions on how to run the code. Submit the code as a zip file in GradeScope under ”Kaggle Code”.

2. Your submission folder should contain your Kaggle notebook named as ”final.ipynb” which would reproduce your predictions exactly. Make sure to fix the random seeds so that the generated predictions are exactly matching your submitted prediction file.

3. 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 could be found in Kaggle website (go to : Overview→ Evaluation).

4. You must submit a written report according to the general layout described above. The report must be submitted in GradeScope under ”Kaggle Report”.

Submission Instructions

For this project, you will submit the report and the code to Gradescope. Make sure we can directly run your notebook in Kaggle. We should be able to run your code without making any modifications. You need to make a team submission (not an individual submission) for both code and report. The competition ends on December 5th and the report is expected by December 8th.

Late Submission Policy

The late submission policy is the same as the default policy used for the other assignments.

Evaluation Criteria

Marks will be attributed based on 40% for performance on the private test set in the competition and 60% for the written report. For the competition, the performance grade will be calculated as follows: the instructors will set up a number of baselines which are each associated with a given grade.

1. Any team performing under the lowest baseline will receive 0%.

2. We have a single baseline which gives you a score of 80%.

3. The highest scoring team will get 100%.

4. All teams will have their grades based on a linear interpolation using their score and the above three thresholds.

For the written report, the evaluation criteria include:

1. Technical soundness of the methodology (pre-processing, feature selection, validation, algorithms, optimization).
2. Technical correctness of the description of the algorithms (may be validated with the submitted code).
3. Meaningful analysis of final and intermediate results.
4. Originality of the approach.
5. Correct insights and analysis on the link between certain attributes and the target.
6. Clarity of descriptions, plots, figures, and tables.
7. 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.