



CMP4040 Project Document

Project Description:

The main objective of this project is to apply machine learning concepts and algorithms to a real-world problem. The selected problem for this semester is "Body Level Classification". First, you need to explore and analyze the given dataset to find the best way to approach it. Then, you have to apply at least three machine learning algorithms (taught in this course) to the dataset in the aim of solving the problem. Moreover, you are advised to explore further methods to enhance the performance.

Problem and Dataset Description

You are required to solve a classification problem for human body level based on some given attributes related to the physical, genetic and habitual conditions. The given attributes are both categorical and continuous. The human body level can be categorized into (4 levels/classes).

You are given 16 attributes and 1477 data samples, where classes are not evenly distributed. Try to build models that can adapt to the class imbalance to achieve the best possible results.

The exact description of each attribute can be found in the README attached with the dataset.

Team Formation

You are required to form teams and submit a single spreadsheet containing the formation of all teams within a week from project announcement.

Deadline: March, 15th, 2023, 23:59.

Project Implementation and Report

You should apply machine learning with the aim of solving the given problem. Try to document your steps from the start as it will help you in writing the report. The workflow should be as follows:

1. Analyze the dataset. The goal is to familiarize yourself with the dataset before diving into the upcoming steps. Try to visualize the dataset, build histograms of features or outputs, look at random samples and look for outliers, etc. It is also useful to use a baseline (e.g., ZeroR) to put your results into context.
2. Apply at least 3 different machine learning methods to the given problem. The methods should be selected from the ones covered in the course. None of the methods are allowed to be deep

learning. For each method, you should test the **effect of the hyperparameters** and try to find a set of hyperparameter values that work well for the problem.

3. You should apply all the concepts you learned in the course.
4. Report your findings in a clear and concise manner, along with comprehensive results analysis. The results analysis and comparison will have a high impact on the final grade.
5. It's important that you try to achieve the **highest performance possible**, as the project is basically a competition and a part of the grade will be allocated to the performance achieved on an unseen test dataset compared to your peers. You can try as many models as you want for this.

Deliverables:

- 1- Project Report (PDF) which should contain:
 - a. The team number and a list of member names (alongside their section and bench numbers).
 - b. The contribution of each team member (This must be clearly stated as each member is graded individually).
 - c. The problem definition, motivation, and evaluation metrics, including links to the dataset and any references. (This part can be copied from the proposal).
 - d. Your results (the dataset analysis results and the experimental results).
 - e. A discussion of your experimental results (an in-depth analysis).
 - f. Your conclusion.
- 2- The project source code.
- 3- Presentation (Slides) summarizing your work and analysis (to be presented during the project discussion).

Deadline: May, 20th, 2023, 23:59.

Rules:

- 1- Each team can consist of up to 4 members.
- 2- Any evidence of plagiarism will result in receiving ZERO points for the project.
- 3- Each member of the team will be graded individually for their contribution and understanding.