# CP468 Artificial Intelligence

# Project Instructions

# Due date: 11:59 PM, August 10, 2023

**General Guidelines**

- The goal of this project is to apply the Artificial Intelligence/ Machine Learning methods described in this course to work with real-life data sets. You will demonstrate an understanding of the machine learning methods covered, their applicability to specific problems, possible pitfalls, and creative methods of combining them to identify data patterns.
- You will work in a group of a maximum of 3 students per group, although you can work independently if you want.
- You can choose any application area of machine learning, including prediction, classification, and clustering.
- You can select any data set described in the "Datasets" section in the instructions.

**Primary Questions**

- To build an Artificial Intelligence/machine learning model for solving a real-life problem using data of your interest. It should take a systematic step approach:
    - Selection of Problem Area:
        - Choose the problem area that interests you the most.
    - Choosing/collecting dataset:
        - You may include the details about (it is just a reference list; you may add the relevant information depending upon the dataset and the problem you have chosen)
    - Performing Data Preparation:
        - Missing values in your data and how will the missing values be dealt with.
        - Any relationships between the variables.
        - How has the dataset been divided into training, validation, and test sets?
        - If you are doing any clustering, are there clusters in the data? How many? How well separated, are they?
    - Model Planning:
        - What classes of algorithms should be applied to problem areas?
    - Model building:
        - Here, the model learns from the data and adjusts its parameters to minimize the error or maximize the performance metric defined.

- Model Evaluation:
  - Apply the build model to new, unseen data/environment to evaluate its real-world performance.

### Deliverables of the Project (The project is worth 20% of your term marks)

The project includes a report and a presentation, which have equal weight.

- Final Report: A final project report should be divided into sections and subsections, including the following components:
  - Title Page
  - Abstract
  - Introduction
  - Project Description
  - Methodology
  - Results
  - Conclusion
  - References (if any) – Please use either 'Harvard' or 'APA' referencing system.
- Presentation: The recorded video presentation must be submitted.
  - Presentation covering the project details, method used, and project findings.
  - Logical partitioning of the work among the team members must be mentioned on one slide.
  - There is no limit on the number of slides.
  - You must follow the time constraint of 10 mins (at most).
  - It should also be structured logically for equitable presentation.
  - The video presentation must be in MP4 format.
- You are responsible for submitting a zip file to the drop box at Mylearningspace containing all (mandatory) of the following files:
  - Project Report (PDF)
  - PowerPoint Presentation (PDF) (that used in video presentation)
  - Video Presentation (.mp4)
  - Code files

## Datasets

For your project, you are going to need data. There are lots of online sources, some of which are listed below. If nothing below grabs you, I'd encourage you to look for yourself. Some general links have lots of data online at the bottom of the list. Some specific datasets include:

- Microarray data: Data link for Golub data has been found. Go to this link "http://statwww.epfl.ch/ davison/teaching/Microarrays/lab/classification.html" and go to the paper "Molecular Classification of Cancer: Class Discovery and Class Prediction by

Gene Expression". The paper name is "Golub et al 1999.pdf". Note that other data here might be applicable - you're welcome to look around.

- Tecator data: The data are available at statlib, at "http://lib.stat.cmu.edu/datasets/tecator". The data consist of 240 samples of meat, which have been analyzed using near infrared absorbance spectra. 100 absorbances at different spectra are recorded for each of 240 samples. Although three potential responses are all continuous, you could analyze a discretized version of one or more of these.
- Thyroid Disease Database (more than 3000 cases, 22 variables): From the UCI machine learning repository, this data gives various medical attributes of patients, and whether they have a hyperthyroid condition. The data are described under the title "Thyroid Disease Database" at http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease"
- US Universities Data (1300 cases, 30 variables): This dataset is taken from the 1995 U.S. News & World Report's Guide to America's Best Colleges. Information on around 30 variables for each of over 1300 American colleges and universities is given. See the link "http://lib.stat.cmu.edu/datasets/colleges/readme" for details. You could predict several variables including public/private.

**Websites listing datasets.**

- Google Dataset Search: Dataset Search lets you find datasets wherever they are hosted, whether it's a publisher's site, a digital library, or an author's web page. It's a phenomenal dataset finder, containing over 25 million datasets.
- Kaggle: Kaggle provides a vast container of datasets, sufficient for the enthusiast to the expert.
- KDD cup datasets (http://www.kdnuggets.com/datasets/kddcup.html) have some interesting data mining problems (see also kdnuggets site below, it has some good links). This includes some drug discovery applications.
- At the University of British Columbia, Dr. William Welch has a long list of drug discovery datasets, (http://stat.ubc.ca/ will/ddd/).
- The Big Bad NLP Database (https://datasets.quantumstat.com/): This cool dataset list contains datasets for various natural language processing tasks, created and curated by

**Evaluation Criteria**

- Presentation of content: The report should be well-organized and written. If figures or tables are in the appendix, references in the text should be explicit. The figures and table must be given captions appropriately.
- Correctness: The diagrams, methods, results etc., should be correct.
- Completeness: To develop a project following the appropriate methods.