Learning Outcome

The Machine Learning project is a venue for students to achieve the learning outcomes below:

LO1. Compare and contrast existing literary works, machine learning techniques, tools, and approaches [CS02.01, CS03.02].
LO2. Develop machine learning solutions after decomposing real world problems [CS02.03, CS05.03].
LO3. Adapt existing algorithms, libraries, tools, and approaches in the development of machine learning solutions [CS03.03, CS05.01].
LO4. Design solutions of varying level of complexities to re-implement existing machine learning applications [CS04.01].
LO5. Write and present the machine learning project pipeline comprehensively including choices, assumptions, and consequences of decisions [CS07.02, CS07.03].

The goal of this project is for the student to be able to perform a Machine Learning analysis given a specific data set. The student should be able to identify patterns, questions from the data set, and perform specific operations that will help them understand the data set. They should also be able to analyze the results. In this approach the student should be able to use existing AI/ML techniques in coming up with a scientific inquiry.

Deliverables

Groups must submit a Jupyter notebook containing the whole pipeline of their methodology. This includes and is not limited to data preprocessing, model building, evaluation (validation and testing), and fine tuning.

The notebooks are expected to be verbose. It should walk the reader on the steps the student made to make their model work. The notebooks must also show the authors' efforts to make their model work.

The notebooks are expected to come with their latest checkpoints (check hidden files of your notebook's directory). Groups must also make sure their notebook will run properly from start to bottom in a single kernel run.

Instructions

1. Form a group composed with maximum of 3 members.
2. Download Anaconda from https://www.anaconda.com/
3. Find and download a dataset of at least 1,000 samples and 5 features from sites like Kaggle and UCI Machine Learning. You may also choose datasets that are not from known sites as long as you clearly describe the dataset. The dataset must be suitable for classification or regression.
   Below is a list of dataset sources:
   - Kaggle
   - Datasets from Center for Empathic Human-Computer Interactions (CEHCI)
   - Datasets from Center for Language Technologies (CeLT)

- Datasets from Center for Automation (CAR)
- University of California at Irvine Machine Learning Repository (http://archive.ucs.uci.edu/ml (Links to an external site.))
- University of Irvine Knowledge Discovery in Databases Archive (http://kdd.ics.uci.edu (Links to an external site.))
- Gait Database (http://physionet.ph.biu.ac.il/physiobank/database/gaitdb (Links to an external site.))

4. Identify a problem statement or question that can be solved or answered by performing Machine Learning tasks on the chosen dataset.
5. Perform Machine Learning tasks so that an analysis can be derived from the said data set. Use Scikit-learn in Python to perform the Machine Learning tasks.
6. Use at least two Machine Learning models, and compare their performances using performance metrics like Confusion Matrices, Accuracy Scores, Precision, Recall and F-measure.
7. Write a documentation describing the whole process and the results using Jupyter notebook. You may use the previous technical notebooks as your guide.

Final Deliverables

The Machine Learning project has two components – a technical notebook and a presentation. Prepare the technical notebook and the presentation containing the following. The documentation must be written on the notebook. You may use the previous technical notebooks as examples.

1. **Introduction**
    - Introduce the problem statement/question.
    - You may start with defining/describing the domain of the problem/question. For example, the problem statement is recognizing the emotion of a person. You may start with defining emotions, what are the different emotions, and how are emotions expressed. You may also give some statistics related to the topic.
    - After the introduction of the domain, discuss the motivation. Why the problem or question must be solved/answered.
    - End the introduction by formally stating the problem statement and the approach. Then, enumerate or explain the possible benefits if the problem is solved.
2. **<Name> Dataset**
    - Describe the dataset
    - Show the demographics of the participant (if applicable)
    - Enumerate and describe the features
    - Enumerate and describe the labels
    - Show the distribution of the classes
3. **Methodology**
    - For every block of code, explain the purpose.
    - Simple data analysis

- Describe your data and show what kind of initial features you are dealing with
- You can point anomalies/outliers in the data
- You may also use data visualizations.

- Data preprocessing/cleaning
  - Explain why the data was preprocessed that way
  - If you removed some data, explain why removing those data was necessary
- Feature extraction
  - Explain (even if briefly) what these features are, and why they may help (if applicable)
- Model training
  - Explain why you chose the algorithms you will use
- Feature selection
  - If you removed some features, explain what method you used to determine which features must be removed.
- Hyperparameter tuning
  - You may use grid/random search for hyperparameter tuning

4. Results and Analysis

- Discuss the performances.
- You can have 3 subsections for this:
  - First subsection for first Machine Learning model. Briefly compare the models created given the different hyperparameters used. Illustrate the comparison using a graph. Give some hypotheses why the performance is decreasing or increasing as you modify the hyperparameters. End with stating the optimal hyperparameters based on the experiments.
  - Second subsection for second Machine Learning model. Do the same as the first subsection.
  - Third subsection for the comparison and analysis of the results using the 2 Machine Learning models. Compare the performances of the best models. Give some hypotheses why one is better than other.

5. Conclusions and Recommendations

- Summarize what you did in 1-2 sentences.
- Briefly discuss the best model and its hyperparameters.
- Give some recommendations to improve the performances.

6. References (Follow the APA format)

7. Appendix A. Contribution of Members

Name	Contributions

Submission Policy:

- Submit the presentation slides and the technical notebook of your Machine Learning project on **30 May, 2021** (2359).
- Late submissions will incur corresponding deductions.
- Submission of the correct files is your responsibility. Please check the files before, and after submission.
- The presentation can be scheduled thru Canvas.
- **Plagiarized works will automatically be given a grade of 0.0 for the course**