WIA1006/WID3006 Machine Learning
Group Assignment (5~6 per pax)
20%
Deadline 3rd June 2024 11.59pm
Submission on SPECTRUM (One submission per team)

This project is designed for FCSIT UM undergraduate Computer Science students who are taking the WIA1006/WID3006 Machine Learning. The submission is in week 13, and 3-5 finalist will be narrowed down to a final pitching in week 14.

1. Themes

For the theme and dataset of the project, your group can choose from the domain below. Please refer to Section 3.0 dataset for further elaboration.

1. Oil and Gas data (Eg: Reservoir data, operational safety data)
2. Utilities data- (Eg: Water, Electricity consumption)
3. Finance, Banking (**Strictly No credit card fraud detection dataset**)
4. Others but must be collected/mined by you. Details, please refer to 3.0 Dataset as below.

2. Machine Learning Flow

In this project, you will be workingin a group of 5-6 people to explore and analyze a specific theme using machine learning techniques. The goal is to train, optimize and compare several machine learning models (min 5 models) and compare their performance in solving real-world problems. The standard Machine Learning pipeline is applied here as follows.

1. **Design your project**: As a group, you should first brainstorm and come up with a specific theme that you are interested in exploring. You are encouraged to refer to Kaggle or some other Githubs and repo for ideas, but the Machine Learning problem that you want to solve must solely come from you.
2. **Data Collection/acquisition/mining**: Once you have decided on a theme, you will need to collect/find data that is relevant to your project**. You are not allowed to use dataset from Kaggle or ones that can be easily access using scikit-learn libraries (such as Iris data, house pricing data, Wine dataset)**
3. **Data Pre-processing**: After collecting your data, you will need to preprocess it to make it suitable for machine learning. This may include cleaning, normalization, and transformation.
4. **Feature selection/Feature extraction**: You may want to do feature selection or extraction using techniques such as PCA here before training.
5. **Model Selection**: Once your data is ready, you will need to choose an appropriate machine learning algorithm for your project. You may choose from supervised or unsupervised learning, and select the model based on the problem you are trying to solve.
6. **Model Training and Hyperparameter tuning**: Once you have selected the model, you will need to train it on yourpreprocessed data and evaluate its performance using appropriate metrics. You may need to fine-tune your model to improve its performance.

7. **Model Evaluation**: Depending on the Machine Learning task you have chosen, what is the best evaluation methods for your models? How do you compare your model against other models? How does your model compare to auto-sklearn? (https://automl.github.io/auto-sklearn/master/)

8. **Final Deliverable and Conclusion**: Finally, you will need to present your findings and results in a clear and concise manner. Please submit in SPECTRUM

   a. Your Google Colab/Kaggle Notebook link. If you are using local notebook, submit 2 files, one as .ipynb file and another one as .pdf file. Please state your team member's names

   b. Your presentation slides. Please state your team member's names

   c. A link to a 5min video presentation of your project. Please state your team member's names

   d. No report or poster is needed.

3. Dataset (VERY IMPORTANT)

In this assignment, **You are not allowed to use dataset from Kaggle or ones that can be easily access using libraries (scikit-learn) such as Iris data, house pricing data, Wine dataset,diabetes data, heart attack data, and others** . Instead, we encourage you to collect, acquire or mine the dataset yourselves. 5 marks will be given to the team that has collected, acquired, or mined a good dataset. Here are the several ways that you can collect, acquire, mine your own dataset.

1. Find the data from Journal/Conference papers. Sometimes the repository (Github) link to the dataset can be found in the journal paper. It may or not be publicly accessible, you may need to email the authors for access.
2. Find the data from open data sources such as data.gov, open.dosm.gov.my
3. Collect the data yourself by doing a survey (physical or online). But if you choose this path, please make sure you have enough samples/respondents and also enough time to do the sampling.
4. Mine online website data using webscraping methods such as BeautifulSoup (refer to WIA1006_Data_Mining.ipynb)
5. Acquire data using API (refer to WIA1006_Data_Mining.ipynb).

4. Assessment

The assessment of the project will be divided into two (2) sections:

1) TECHNICAL KNOWLEDGE
   a) Relevance and Significance of the Problem (10 points)
   b) Data Collection and Preprocessing (20 points)
   c) Model Selection and Performance (30 points)

2) SOFT SKILL EVALUATION
   a) Presentation Quality (20 points)
   b) Creativity and Innovation (10 points)
   c) Teamwork and Collaboration (10 points)

This assignment shall evaluate soft skills elements:

A. COMMUNICATION SKILLS (CS1, CS2, CS3) *

PRESENTATION (Flow of discussion, Presentation (language/fluency/idea coherency), Teamwork, Effort and Q&A skills).

B. CRITICAL THINKING AND PROBLEM SOLVING (CT1, CT2, CT3)
Creative and critical thinking (Innovation, connecting, synthesizing knowledge and transforming ideas into new forms/solutions) for Individual Project.

C. MORAL AND PROFESSIONAL ETHICS (EM1, EM2)
Moral & Professional Ethics – ethics in presentation of results in online/live presentations.

Total Points: 100 (20%)

Deadline: 3rd June 2024 by 11.59 p.m.
You are required to submit your recorded presentation including the project demo, coding and PowerPoint slides. The duration of the presentation is capped to 5 minutes. Each of the members needs to present their technical part contributing in this project.