UNIVERSITY OF JAFFNA
FACULTY OF ENGINEERING
EC 9560: DATA MINING
INDIVIDUAL PROJECT - SEPTEMBER 2022

The purpose of this project is to understand how to train a supervised machine learning model, how to tune its parameters, how to evaluate the model using techniques such as cross validation, and how to make predictions of an unlabeled test dataset.
In this project students are encouraged to learn new classifiers and techniques to get better prediction for the unlabeled test set.

- Visit the website **analyticsvidhya**
- They have active and closed challenges.
- Choose one challenge. No problem whether it is active or closed.
- Note that if more than one groups are choosing same heading, make sure there are no copies.
- If you are choosing the active challenges and you can finish it before the deadline, you can be a participant.
- You are expected to submit the heading and the group detail on or before **10.10.2022.**

Evaluations

- This covers your lab hours (12 hours) and assignment hours (15 hours).
- You are supposed to attend all the labs to do this project and submit the lab reports. It will be marked for 10% of your final marks.
- Further you need to submit two reports: one during mid (progress report) and another during final (end report) –for assignments – 30% of your final marks.
    - Mid report (progress): Deadline 7th of November.
    - End report: Deadline 5th of January.

Your end report should cover the following,

1. Brief description
    (a) about the dataset – how many attributes, how many classes, how many instances in each class, etc. [20 marks]

(b) regarding whether it is a classification, regression or a clustering problem. [10 marks]

(c) of the pre-processing techniques (such as handling missing values, outlier removal, feature scaling/normalization, etc.) that you used (if any). [20 marks]

(d) of the machine learning method that you used for this problem and the reason for selecting that method. You may use any machine learning approach (e.g. Boosting, Random Forest, Neural Nets, SVM, etc.), but a basic understanding of the approach that you used is required. [20 marks]

2. Explain the evaluation criteria you have used to validate the model. [10 marks]

3. Briefly explain the parameters of the model and explain how did you tune the parameters (e.g. cross validation)? [10 marks]

4. Explain the meaning and the purpose of cross-validation, and report 5-fold cross validation results on the training set (e.g. in the form of accuracies or confusion matrix for classification problems or mean squared error or sum of absolute error for regression problem). [10 marks]

Note:

- This project contributes 40% of your final marks of the module.
- You are free to use any library for this. If needed you can convert the given files in to the format needed by the library.
- Your report may include explanations, diagrams and/or screenshots of the results.
- Please submit your reports in **pdf** format. The filename should be your registration number.