**University of Canberra  
Faculty of Science and Technology**

**Pattern Recognition and Machine Learning**

**Project Template for Computational PRML Project**

**Duration: Weeks 3-13, Semester 2, 2023**

**Project Title:**

**Group No.:**

**Project Team Members (maximum of 2):**

**Project Summary:** From 1-2 research papers or relevant problems in the PRML areas with accessible datasets identify a project of your interest. This will form your Unit Project for the full semester. You may use keywords on topics to search for relevant papers and pattern recognition techniques/algorithms on google or google scholar or other community sites such as Kaggle or GitHub. Staged through the semester, you will develop the problem identified, develop, and evaluate a model built from extracted patterns from the selected dataset using the best performing algorithm from 3-4 algorithms you have chosen from lectures and/or literature survey/online resources.

Develop a report of a maximum of 2 pages for the first stage of this project as Part A of Assignment 1 covering the following first 4 steps.

The project i). could be an application of PRML to a real-world problem with a given (accessible) dataset; or ii). could be an application reproducing an existing application with some improvements (based on any gaps/deficiencies reported in previous application or research); or iii). could be a PRML related research problem with new research outcomes (building on recent reported research in literature). Any PRML-based project would suit this exercise as long as it broadly meets the characteristics of a PRML project including Modelling and problem-solving and/or PRML-based research characterisation. The scope of the project will need to be adjusted so it is doable within the semester.

**Stage 1 – Assignment 1, first 3 Steps (A1-A3) below**

A1. Identify & define your PRML Project. Explain why it fits the characteristics of ‘pattern recognition’. Give a motivation (such as why a computational model is needed) and give some context to why a PRML solution is required. How can the model be used for problem solving and decision-making?

**Project Suggestions**For this task, you need to select one suitable topic from following options or any other PRML problem with similar complexity of your interest but with a dataset accessible to you.

1. Infectious disease modelling – COVID-19 pandemic, SARS or other pandemics – transmission patterns, identifying hot-spots, different forms of transmission, peaking times, predictions. You could develop a model learning from data to better manage any future such cases of pandemics in the domain of your investigation.
2. Breast cancer, diabetes, skin cancer recognition, modelling effect of medication and resistance to cancer cases etc (UCI datasets – at https://archive.ics.uci.edu/)
3. Mental health modelling – depression, stress, anxiety – detection, classification, and possible remedies to better manage.
4. Handwriting Recognition
5. Document classification
6. Voice recognition – accent, language, origin detection – as a biometric signature
7. Fingerprint classification – as a biometric measure
8. Detection of individuals from face images
9. Pattern recognition for insights from brain signals (Electroencephalogram – EEG): Person recognition, Age & Gender recognition, Emotion recognition, Epileptic Seizure detection.
10. Modeling any behavioral pattern – lie detector, affective measure (smile, anger, sombre, other moods)
11. Modeling vision impairment – image detection measures
12. Number plate detection - for tracking car movements across Local Government motivated by modelling mobility in affected regions to better manage COVID (sample datasets available online)
13. Detection of plants by uploaded images through mobile cameras (sample datasets at UCI and online)
14. Cancer checks through analysing uploaded images of skin lesions (<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T>)
15. Traffic modelling in ACT (ACT Gov data site: <https://www.data.act.gov.au/Transport/Traffic-Links-Stats/jn4p-azhb>)
16. Recognising baby behaviours through sound, facial expressions - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4140581/>.
17. Speech to text converters, in english-to-english <or any language to any language>
18. Bush fire prediction/detection/modelling based on patterns.
19. Weather forecasting
20. Several other ideas also at <https://analyticsindiamag.com/20-machine-learning-datasets-project-ideas/> or GitHub and Kaggle.

**Datasets**You need to select one suitable dataset from the following options or some dataset publicly accessible for your selected project.

1. UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets.php>
2. Australian Government Data: <https://data.gov.au.>
3. Kaggle: <https://www.kaggle.com/data>
4. Amazon: <https://registry.opendata.aws/>
5. Google: <https://cloud.google.com/bigquery/public-data/>
6. Electroencephalogram (EEG): <https://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html>

**Learning Models**  
You may choose any 3-4 of the following learning algorithms or any others that may be suited for your problem either from the material covered in the unit or from other resource material

* Linear Regression
* Logistic Regressions
* Naïve Bayesian
* K-Nearest Neighbour
* Decision Tree
* Support Vector Machine
* Convolutional Neural Networks
* Clustering
* Ensemble models
* etc

A2. Investigate and characterize the problem in order to better understand the goals of the project. Identify what questions will be investigated in the project and how the models learned will help in answering the questions. How will the model be validated and used to predict unseen problem cases?

A3. Explain why the proposed project qualifies as a pattern recognition and machine learning problem

**A4. Write and submit a report for Assignment 1: Part A responding to each of the sections A1-A3 above. (Max of 2 pages)**

**Draft outline for future stages:**

**Stage 2: Assignment 2 (Part A)**

B1. Select a dataset for your project – you may use the 10 times rule to just check on the size of your data! The rule is if you have x number of parameters to model you should have at least 10x data records to train. For example if you want to classify cats and dogs based on 1000 parameters, you will need 10,000 pictures to train the model.

B2. Analyze Data: Use descriptive statistics and visualization to better understand the dataset.

B3. Prepare Data: Use data transforms/scaling/re-shaping and feature selection in order to reduce dimensionality and better expose the structure of the prediction problem to the modeling algorithms. Summarise your data.

B4. Decide which algorithms will be suited for your project. Explain your reasoning. Select 3-4 algorithms for spot-check and determine the best based on accuracy evaluations.

**B5. Write and submit a report for Assignment 2: Part A responding to each of the sections B1-B4 above. (Max of 2 pages)**

**Stage 3: Assignment 3**

C1. Evaluate Algorithms: Divide data for learning and test cases (you may try some splits eg 70-30 or 80-20 and investigate any accuracy issues). Design a test harness to evaluate the selected algorithms on the chosen data and compare selecting with explanation the top 2 to investigate further. Use cross-validation to check on overfitting and whether you require approaches to deal with overfitting (as discussed in lectures).

C2. Improve Results: Use algorithm tuning (using parameter tuning/hyper-parameter tuning) and possible ensemble methods to further improve the selected well-performing algorithms on your data. What methods would you use to fine-tune and optimise your results? This may be tuning internal model parameters or hyper-parameters, using appropriate solvers & determining the best parameters using gridsearch etc.

C3. Present Results: Finalize the outcomes from the chosen learning models and determine the best performing model from the classification report.

C4. Show outcomes visualising test cases against training cases for the best performing model.

C5. Make predictions using your optimised model for sample unseen cases and present results.C6. Create a standalone model on entire training dataset

C7. Appraise any ethical and privacy issues emerging from the problem and how you have addressed them

C8. Save model for later use – explain how this saved model can be further developed with new data and how it can be used to classify/predict outcomes for unknown cases.

**Final Report for the Unit Project – final project submission (including Stage 3 work)**

**The final report will include your responses to C1-C8 above.** Write your report integrating all the outcomes from each of the above 3 stages. The report has to be a technical paper with author details, an abstract and the contents organised to cover the 3 stages of your work. I should be at most 6 pages (in single-column format, including references) with a font size between 11-12 point. It should be in IEEE format in A4 format – as specified here: https://www.ieee.org/conferences/publishing/templates.html. Save it in a file called PRMLProjReport-GpNoxxx.pdf. Save your code in either a .py or .ipynb file with in-code documentation. Zip the python file to Code-GpNoxx. Submit both the report and the zipped code through Canvas by the due date.

Penalties will apply if the report does not satisfy the requirements. Remember to clearly cite any sources (including books, research papers, course notes, code etc.) that you used/referred to while designing aspects of your solution for the project.

**Referencing guidelines**   
Use the IEEE referencing style for the reports. You must acknowledge all the sources of information you have used in your assignment and cite as references.