University of Lincoln
School of Computer Science
Assessment 02 Briefing 2023-2024

The use of AI tools to generate all or part of your assessment submission is **not** permitted unless specifically mentioned below.

Module Code and Title: CMP3751M Machine Learning
Contribution to Final Module Mark: **60%**
Description of Assessment Task and Purpose: This is Assessment 2 and is an **individual** assignment.   For this assessment, you are required to implement a pipeline for processing a machine learning dataset. You will be required to try out different machine learning approaches, implement an appropriate evaluation strategy, and compare and select between different models. You will also be required to analyse the outputs of your trained models, taking into account the data you are working with.   You are required to **download and modify** the Jupyter notebook "ML_2324_assessment.ipynb" provided for the assessment, by **implementing your solution** (in notebook cells marked as **SOLUTION CELL)** and **analysing the obtained outputs** (in notebook cells marked as **ANSWER CELL).** You **must not modify any other cells** in the provided notebook file. You are required to follow the implementation structure (i.e. use the function, class and method names detailed in the notebook, and follow the required return format of any function and methods). The 'batteries.csv' file containing the dataset is also provided and necessary to run the provided notebook. Make sure to fix all the random seeds in any parts of your solution, so it can be reproduced exactly. The notebook, as provided, runs without errors (without solving the assessment). Make sure that the solution, or the partial solution, you hand in, also runs without errors on the data provided. If you have a partial solution causing errors which you would like to show, please include it as a comment.   The assessment is divided into six different sections, where **each section has a programming and a discussion component.** The total weighting of the implementation component is 55% of the assessment, and the discussion 45% (please check the CRG for details). **All discussion must be supported and evidenced by your implementation in order to earn marks. Sections progress roughly in order of increasing difficulty, except the last section ("Final model evaluation") which should be attempted by everybody:**   **1. Loading the dataset**   You are provided a dataset describing the physical and chemical properties of the Li-ion batteries, classified on the basis of their crystal system. While this is a real dataset, please take care that you will be working with a **modified and reduced version** of this dataset, so you will need to use the *batteries.csv* file provided for the assessment, rather than downloading the original dataset (which is referenced in the brief).   You are required to load this dataset into *numpy.array-s,* ensure that the class label is encoded as an *integer*, and properly handle any missing or outlier values.

You are also required to describe the dataset, in terms of number of samples and number of classes, and justify your approach to handling missing or outlier values.

2. Simple classification model

To get the feel for the dataset, the first step will be to build train a simple classification model for this dataset. First, you will need to set aside 20% of the data for testing, and use the remaining 80% to train your model. Then, you will be required to train a simple classifier, (of your choosing) with fixed parameters on the dataset, and calculate accuracy on the test set. For calculating the accuracy, you will be required to implement a function called *model_accuracy(y_test, y_pred)* (full function specifications are provided in the Jupyter Notebook).

You are also required to discuss the advantages and shortcomings of the evaluation strategy implemented through this task, in terms of both the *data split* used for evaluation and the *choice of evaluation metric.* Based on the information about the dataset you have so far, you are also asked to commend on the performance of the model you have trained for this task.

3. Improved evaluation strategy

Based on the shortcomings of the simple evaluation strategy you have identified in the previous task, for this task you are required to propose a better evaluation strategy. Make sure your chosen strategy uses all the samples in the dataset to report the result. You are required to implement a function *evaluate_model(model, X, y)* to carry out your proposed evaluation strategy (full function specifications are provided in the Jupyter Notebook).

You are then asked to discuss your chosen evaluation strategy, including both the data split and the evaluation metrics.

4. Different models and parameter search

In this task, you will use your implemented evaluation strategy to compare different machine learning models of your choosing. Fit at least three different (types of) machine learning models to the provided dataset, where at least 2 out of your 3 chosen types have different model parameters which can be adjusted. Try different parameters for all of your models (which have parameters). Use a single summative metric of your choice to choose between the different types of models, and the models with different parameters. Finally, choose the best model of each type according to your proposed evaluation strategy.

You are then asked to discuss your choice of models, and your procedure for adjusting the model parameters. Discuss how you reached the decision about the best model amongst the models of the same type (which metric was selected, and why). Also discuss any shortcomings of your approach and how (and if) you could improve on this. After evaluating these models on the dataset, discuss and compare their performance on the provided data.

5. Ensembles

Combining different weak classification models can improve the overall performance of the model. Implement bagging for each of your three classification models chosen in the previous task, using any additional evaluation you require for created bagged ensemble models.

You are provided with code which will run (your own) evaluation procedure on the three created bagged ensemble models. You are also provided with the code which will combine your three base models (from Section 4), and your 3 bagged models (from this Section), and evaluate these voting-based ensembles.

Discuss the effect on bagging on your base models. Discuss how you chose the bagging parameters, and justify your choice. Discuss the effect using the voting ensemble had on your model performance. Compare the effect of a voting ensemble on the ensemble models to the effect on the base models.

6. Final model evaluation

Try to engage with this Section even if you have not completed all the sections above.

Based on all the experiments performed for this assessment, choose a single best model amongts the ones you've trained, evaluate it with your evaluation procedure and also display the confusion matrix.

Discuss the performance achieved by this model.

Use of programme code from libraries and external sources:

You **are allowed to use any functionality provided by *numpy, sklearn, pandas,* and *scipy*** packages (and any other packages available by default on the University of Lincoln lab machines) **except sklearn.ensemble.RandomForestClassifier.** You **are allowed to use code from external sources,** however any such code needs to be clearly marked both in comments in the code and in the *References* section of your assessment Jupyter notebook. Examples of how to properly reference your sources in a Jupyter notebook are provided within the assessment notebook itself. **The failure to reference the external sources you have used will be treated as plagiarism under the University of Lincoln regulations.** (Please see below for more information about dishonesty, plagiarism and the use of AI tools).

Please see the Criterion Reference Grid for details of how the work will be graded.

Learning Outcomes Assessed:

- [LO2] Using a non-trivial dataset, plan, execute and evaluate significant experimental investigations using multiple machine learning strategies

Knowledge & Skills Assessed:

Subject Specific Knowledge, Skills and Understanding:

Comparing different machine learning approaches, applying different model evaluation techniques, analysing and understanding model outputs.

Professional Graduate Skills:

Creativity, critical thinking, problem solving, effective time management.

Emotional Intelligence:

Self-management

Career-focused Skills:

Professional code of conduct understanding, Project planning, Reflective practice.

Assessment Submission Instructions:

The deadline for submission of this work is included in the school submission dates on Blackboard. Your solution should be created by modifying the provided template ipynb file.

Your solution ipynb should be **renamed** to "ML_2324_ssessment_xxxx.ipynb" *where xxxx is your student number*, and uploaded to Blackboard directly under "Assessment 2 upload".

Date for Return of Feedback:

Please see the School assessment dates spreadsheet.

Format for Assessment:

The submitted solution should be a single .ipynb file written in Python, **renamed** to "ML_2324_assessment_xxxx.ipynb" *where xxxx is your student number*, and uploaded to Blackboard directly under "Assessment 2 upload"

Feedback Format:

Written feedback will be provided via Blackboard.

Additional Information for Completion of Assessment:

This assessment is an individually assessed component. Your work must be presented according to the School of Computer Science guidelines for the presentation of assessed written work.

Please make sure you have a clear understanding of the grading principles for this component as detailed in the accompanying Criterion Reference Grid.

If you are unsure about any aspect of this assessment component, please seek advice from a member of the delivery team.

Assessment Support Information:

Staff are available via email for simple queries, and during their office hours for more detailed questions, and can provide feedback during this time outside of module hours.

Important Information on Dishonesty, Plagiarism and AI Tools:

University of Lincoln Regulations define plagiarism as 'the passing off of another person's thoughts, ideas, writings or images as one's own...Examples of plagiarism include the unacknowledged use of another person's material whether in original or summary form. Plagiarism also includes the copying of another student's work'. Plagiarism is a serious offence and is treated by the University as a form of academic dishonesty.

Please note, if you use AI tools in the production of assessment work **where it is not permitted**, then it will be classed as an academic offence and treated by the University as a form of academic dishonesty.

Students are directed to the University Regulations for details of the procedures and penalties involved.

For further information, see www.plagiarism.org