UNIVERSITY OF LINCOLN

Assessment Item 2

Module Code & Title: CMP3751M Machine Learning

Contribution to Final Module Mark: 70%

Description of Assessment Task and Purpose:

For this assessment, you are required to implement two different classification approaches: the **k Nearest Neighbours** classifier and the **Decision Tree** classifier. While many machine learning toolboxes offer a range of classifier implementations which are available for use, implementing an approach from scratch and applying it to a real dataset can provide invaluable insight into the workings of different approaches, and help you better identify and understand strengths and weaknesses of those approaches, and when it is appropriate to apply them.

You will also be handling two different types of data, represented by two different datasets, for this task:

- Crystal System Properties for Li-ion batteries dataset ('batteries.csv'): This dataset contains **only numerical values.** The goal of the dataset is to predict the class of a battery on the basis of its crystal system (monoclinic, orthorhombic or triclinic). The sample features correspond to the physical and chemical properties of the batteries. Each sample belongs to one of three calsses, on the basis of the batteries' crystal system. The batteries data used for the assessment is adapted from the full dataset available here.
- Forest Cover Type dataset ('forestcover.csv'): This dataset contains a **mix of numerical and categorical values**. The goal of this dataset is to predict the forest cover type for 30x30 meter cells, based on cartographic variables only (i.e. not relying on any remotely sensed imaging data). The sample features are derived from US Geological Survey data, and collected from four different wilderness areas representing forests with minimal human-caused disturbances. The dataset therefore also includes, for each sample, the information about the geographical location (wilderness area) and the soil type present at the location (one of 40 categories). The samples are divided into 7 classes corresponding to different forest cover types. The forest cover data used for the assessment is adapted from the full dataset available here.

Note: For your work on the assessment, the provided '.csv' files contain only a portion of each of the dataset (specifically, half of the samples). These are in the same format as, and can be used interchangeably with, the final '.csv' files which will be used to evaluate your solutions.

You are required to **download and modify** the Jupyter notebook "ML_assessment.ipynb" provided for the assessment, by **implementing your solutions** or **answering questions** in the indicated notebook cells (marked as **SOLUTION CELL**). You **must not modify any other cells** in the provided notebook file (especially the ones marked as **TESTING CELL**). You are required to follow the implementation structure (i.e. use the function, class and method names detailed in the notebook, and follow the given return format for any functions and methods). The two '.csv' files, containing the datasets (described above), are also provided and necessary to run the provided notebook.

You can find the detailed instructions for each of the tasks in the provided notebook, followed by a **SOLUTION CELL** (where you are expected to write your implementation) and then a **TESTING CELL** (which provides some test cases, and can be used to insure your implementation will process the data provided. **Note:** the testing cells do not test for the **correctness** of your implementation, just that it runs with the provided data). The assessment is

divided into 4 different tasks:

1. Dataset statistics (10%)
2. k Nearest Neigbours implementation (30%)
3. Decision Tree implementation (40%)
4. Model evaluation and analysis (20%)

Please read the full set of instructions and explanations for each task provided in the Jupyter notebook before reaching out to seek clarifications.

You **are allowed** to use any functionality provided by numpy, sklearn and pandas packages (and any other Python packages available by default on the machines in the University computer labs). However, please note that **no corresponding sklearn implementation exists for the handling of categorical data,** therefore achieving the highest marks at the assessment will only be possible by providing your own implementation of the required classifiers.

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:

- Knowledge of different machine learning approaches.
- Understanding different feature types (numerical vs categorical).
- Understand and apply different model evaluation techniques, and analyse their outputs.

Professional graduate skills:

- Analytical skills
- Critical thinking
- Problem solving
- Time management

Emotional intelligence skills:

- Self-management
- Motivation

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_assessment_xxxx.ipynb” *where xxxx is your student number*, compressed into a zip archive and submitted via the appropriate Supporting Documents upload section as a single file. Please remove any other files from the zip archive, including the provided dataset, before submitting.

Format for Assessment:

The submitted work should comprise of an ipynb file written in Python, **renamed** from “ML_assessment.ipynb” to “ML_assessment_xxxx.ipynb” *where xxxx is your student number*. This needs to be compressed into a zip archive and submitted via the appropriate Supporting Documents upload section as a single file. Please remove any other files from the zip archive, including the provided datasets and this briefing document, before submitting.

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 during their office hours and can provide feedback during this time outside of module hours.

Important Information on Dishonesty & Plagiarism:

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

For further information, see www.plagiarism.org

UNIVERSITY OF
LINCOLN