DATA MINING AND MACHINE LEARNING I

National College of Ireland

MSc Data Analytics – Year 1 (MSCDAD_JAN24C_I)

MSc Data Analytics – Year 1 (MSCDAD_JAN24B_I)

MSc Data Analytics – Year 1 (MSCDAD_JAN24A_I)

PGD in Data Analytics – Year 1 (PGDDA_SEP23)

This Terminal Assessment is a replacement for the Spring Exam Assessment.

TABA Release date: 6th March (week 07) 2024

Deadline for Online Moodle Submission – Friday, 5th May 2024 (23:55 Irish Time)

Learning Objective

This assessment examines the following learning objectives for the module:

LO1	Critically analyze fundamental data mining and knowledge discovery methodologies in order to assess best practice guidance when applied to data mining problems in specific contexts
LO2	Extract, transform, explore, and clean data in preparation for data mining and machine learning.
LO3	Build and evaluate data mining and machine learning models on various datasets and problem domains.
LO4	Extract, interpret, and evaluate information and knowledge from various datasets.
LO5	Critically review current data mining research and assess research methods applied in the field

Answer ALL Questions.

Source code (one for each type of machine learning technique) corresponding to the following description must be submitted online on Moodle as a 1 zip file. The report, following the format discussed below, should also be submitted including in that 1 zip file. You also need to fill out three online excel sheets to insert your data: see the description below.

This is a terminal assessment. NO late submission or extension will be allowed.

Turnitin tool will be used to check for plagiarism and similarity for all submitted documents. Submissions with a high level of plagiarism will be referred to Academic Honesty & Integrity committee.

1 Portfolio Submission Overview

You need to create a body of knowledge regarding the comparison of the effectiveness of **TWO** (technically FOUR: see description below) machine learning techniques of your choice using **THREE** sizeable datasets. As a result, you will develop a portfolio of methods that can reveal insights into machine learning methods' performance and application limitations in different contexts.

As a whole, you need to address the following **TWO** research questions:

1. **RQ1 –** Do the different implementations of identically-named machine learning techniques perform exactly the same? If not, what are the outstanding implementations of the identically-named machine learning techniques for specific empirical designs, and evaluation measures? That is, if different implementations of the identically-named techniques perform differently, which implementation is better than the others in each dataset? This will identify the best-performing implementation of each.
2. **RQ2 –** How do the datasets employed in this work differ from each other? Specifically, how the three datasets employed are different from each other in terms of their characteristics which can impact the effectiveness of machine learning techniques?

Now following sections detail the method, you will adopt to address these research questions. Later sections will also describe the format of the report you need to submit with your algorithms and finally the rubric that will be employed to assess this TABA will be presented in the last section.

2 Method

Each part of the project will be focusing on solving some problems related to machine learning and-or data mining. Hence, each part of the research question will be focusing on some particular concern related to these fields. Following we detail each concern one by one:

1. ***Identification of Identically Named ML Techniques***: The first core concern in addressing RQ1 is to identify the identically named ML techniques which can be compared. Towards the identification of such techniques, you can either 1) use the two implementations of the identically-name technique from two different languages, OR 2) you can employ two implementations of the identically-name technique from two different libraries of the same language. FOR EXAMPLE, for kNN machine learning technique, you may compare 'sklearn' python library implementation (https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html) with 'class' library implementation in R (https://docs.google.com/viewer?url=https%3A%2F%2Fcran.r-project.org%2Fweb%2Fpackages%2Fclass%2Fclass.pdf); OR you can compare 'class' library implementation in R with tidymodels library implementation (https://parsnip.tidymodels.org/reference/nearest_neighbor.html) in R. **Note** that you will implement to ML techniques which will result in comparing four different implementations.
2. ***Identifying the Better Performing Technique***: No matter whether the implementations of the identically-name techniques will perform similarly or differently, you need to evaluate the results of your techniques' implementations employing the **HOMOGENOUS EMPIRICAL DESIGN**. For example, your evaluation will require employing some evaluation metrics and the selection of those evaluation metrics will depend upon the types of problem you are solving. That is, if you are solving a classification problem, then you will be employing classification-related metrics (e.g., F1-score, accuracy, etc.) and if you are solving a regression problem you will be employing other metrics e.g., R^2. **You must employ 1) the same pre-processing steps, 2) the same settings of both implementations, 3) the same evaluation metrics, and 4) the same post-processing steps (if required) to compare the two implementations for each of both ML techniques**. The best practice towards this end is to save your data (e.g., in .csv files) after pre-processing and then load that same data to provide input to two implementations for execution. You must also ensure the same setting for each implementation. For example, if you are using kNN classifier k value, train-test split (e.g., 80/20), distance measure (e.g., Euclidean), etc. should be consistent for each implementation. Towards metrics, if you are solving a classification problem you must provide the

F1-score and accuracy. Whereas if you are solving a regression problem you need to provide the results in form of R^2, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

3. ***Identification of Dataset Characteristics which May Impact the Performance of ML Techniques***: The third core concern is how we can characterize the three datasets we will employ towards the comparison of identically named ML techniques. Toward this end, you will assess your datasets for the following seven characteristics:
   - ***I.*** Number of independent and dependent variables
   - ***II.*** Number of records
   - ***III.*** Data types of combination-> Binary, nominal, categorical, textual, numerical
   - ***IV.*** Summary of each variable: min, max, mean, median, and quartiles
   - ***V.*** Data Cleaning:
     - a. Number and proportion of irrelevant predictive/independent variables removed;
     - b. Number and proportion of duplications removed (if any in data) and technique employed towards duplication removing;
     - c. Dimensionality reduction based on PCA/OLS and self-observation;
     - d. Number and proportion of missing values **in total** and number of missing values dealt employing a technique to deal with missing value of your choice;
     - e. Number and proportion of outliers filtered
     - f. First four characteristics of datasets after performing (1-4) data cleaning steps.
   - ***VI.*** Data Normalization: Number and the proportion of total data instances which are normalized and technique used for normalization.
   - ***VII.*** Data balancing characteristics and splitting: You need to provide the number of records in each class. For example, if it's a binary classification problem then how many records in the full dataset do we have for class 1 and class 2? You must use 70% data for training and 30% remaining data for testing. You need to provide a number of records in your training and testing data.

3 Data Collection and Submission Format

Dataset and Techniques Selection: To select three datasets and two techniques, you should not be selecting the same datasets and technique combination as any other student has selected already. This means you can select the same dataset(s) as others have selected ONLY if you are working on different techniques. Similarly, you can select the same techniques as any other student has selected ONLY if you are selecting different datasets as compared to that student. To make sure that every student is selecting a different combination of datasets and techniques, each student needs to fill a row on the following file (Data_Collection.xlsx) and remember your row number as **IDENTIFIER** to insert results (see below) in corresponding identifier row number in results file.

Format of Submission:

1. Use your row number in the above file and fill in the following file corresponding to that number:
   - a. You need to use the *Results_format.xls* file (Results_format.xlsx) to insert your final results of each implementation (that is, F1-score and accuracy if you are solving classification problem, likewise for four metrics (see Section 2 bullet point 2) of regression if you are solving other regression type problem).
2. Create a zip file that will contain 4 implementations of two techniques, 3 datasets you selected, finally pre-processed data provided as input to ML techniques (in csv format), intermediate implementations' results file (e.g., algo predictions stored as csv), and a .doc file that summarizing following:
   - a. In .doc file describe the problem you are trying to address (e.g., Movie Recommendations with Movielens Dataset, Sales Forecasting with Walmart, Stock Price Predictions, Human Activity Recognition with Smartphones, Wine Quality Predictions, Breast Cancer Prediction, etc.)
   - b. Fill in the 7 characteristics of each of the three datasets you employed in tabular format.
   - c. Answers of the following questions in the.doc file where you described your problem:

Q1: Do the two-implementation of identically named technique perform differently or the same?
Q2: If they are performing differently, then what could be the reason? For example, one possible reason maybe they are internally using different algorithms, or implicitly employing some data processing (confirm using the documentation) or maybe some other reason.

3. **You must submit a video presentation that covering ALL the rubric criteria below. You will record your program executing for each of the criteria. You may get marks only for the criteria covered in your video. Your presentation should be preferably 7 minutes long and no more than 9 minutes long.**

4 Assessment Rubric

The assessment rubric is as followed:

High-level Concerns to Address	Explanation and-or Further Segregation	Total of Marks
Completion of Tasks	Technique 1 with 3 Datasets – 20% ■ Datasets Characterization 10% - Characteristics 1: 1 out 10% - Characteristics 2: 1 out 10% - Characteristics 3: 2 out 10% - Characteristics 4: 2 out 10% - Characteristics 5: 2 out 10% - Characteristics 6: 2 out 10% - Characteristics 7: 2 out 10% ■ Technique 1 Modelling 10% - Data Preparation: 5 out 10% - Training/Testing and Results Calculation: 5 out 10%	40%
	Technique 2 with 3 Datasets – 20% ■ Datasets Characterization 10% - Characteristics 1: 1 out 10% - Characteristics 2: 1 out 10% - Characteristics 3: 2 out 10% - Characteristics 4: 2 out 10% - Characteristics 5: 2 out 10% - Characteristics 6: 2 out 10% - Characteristics 7: 2 out 10% ■ Technique 2 Modelling 10% - Data Preparation: 4 out 10% - Training/Testing and Results Calculation: 4 out 10% - Correct answers to questions in Section 3: 2 out of 10%	
Reproducibility of Results	Results are reproducing as it is reported on the teacher machine	30%
Consistent Empirical Design	All pre/post data modeling steps, implementation settings, and evaluation metrics are consistent across both implementations for each of both ML techniques	30%

The proportion of the marks you will get from *reproducibility* and *consistent empirical design* section depends upon your *completion of tasks*. In case you complete half tasks, you will only get marks out of half for reproducibility and consistent design. For example, IF you are only comparing **ONE** technique (two different implementations of one identically named technique) and not **TWO,** then you will get 20 marks out of 40 for completion of tasks, IF the results produced by that implementations’ comparison are reproducible on teachers’ machine then you will get 15 out of 30 for reproducibility otherwise zero, IF you employed consistent empirical design when comparing that technique you will get 15 out 30 for consistent empirical design otherwise zero. Similarly, if you reduce the datasets needed to employ for the 7 characteristics your marks will be reduced similarly.