MCAST

DOC_174_REV_C_ASSESMENT AND IV FRONT SHEET_INDIVIDUAL CRITERIA_CLASSTER.docx

ASSESSMENT AND INTERNAL VERIFICATION FRONT SHEET (Individual Criteria)

(Note: This version is to be used for an assignment brief issued to students via Classter)

Course Title	**B.Sc. (Hons.) Software Development**	**Lecturer Name & Surname**	**Frankie Inguanez** **Alan Gatt**		
Unit Number & Title	ITSFT-606-1618 | Applied Computational Intelligence				
Assignment Number, Title / Type	2, Classification and Reporting / Home				
Date Set	01/04/2024	**Deadline Date**	28/04/2025		
Student Name	Lee Xerri	**ID Number**	446203L	**Class / Group**	SWD-6.3A

Assessment Criteria	**Maximum Mark**
KU1.3 : Outline different forecasting techniques.	5
AA3.1 : Complete a forecasting report using different techniques.	7
AA2.1: Analyse a given data set and identify data cleaning issues.	7
AA2.4: Arrange data in preparation for use within a classifier algorithm.	7
KU2.5: Assess the outcome of a classifier by plotting the appropriate graphs and interpreting them.	5
AA3.3: Apply various techniques to determine the statistical relevance of fields in a dataset.	7
SE3.4: Design and develop a classifier for a given data set.	10
AA4.1: Illustrate and analyse various statistical graphs documenting an exploratory analysis of data.	7
KU4.2: Explain decisions taken by providing supporting evidence in a report.	5
SE4.3: Recommend a decision to be taken based on a documented analysis.	10
SE4.4: Express findings and results in a presentation.	10
Total Mark	80

Notes to Students:

- This assignment brief has been approved and released by the Internal Verifier through Classter.
- Assessment marks and feedback by the lecturer will be available online via Classter (Http://mcast.classter.com) following release by the Internal Verifier
- Students submitting their assignment on Moodle/Turnitin will be requested to confirm online the following statements:

Student's declaration prior to handing-in of assignment

✞ I certify that the work submitted for this assignment is my own and that I have read and understood the respective Plagiarism Policy

Student's declaration on assessment special arrangements

✞ I certify that adequate support was given to me during the assignment through the Institute and/or the Inclusive Education Unit.

✞ I declare that I refused the special support offered by the Institute.

Instructions to Students

- This assignment carries a total of 80% from the final module mark.
- You are requested to upload all content in a zip file on Moodle. Suggested content is the project dataset as CSV and 1 Python Jupyter Notebook.
- Copying is strictly prohibited, and any students caught will be subject to the respective MCAST Disciplinary Procedures.

Task 01: Regression

With sustainability being a global priority, nations are tracking the amount of waste generated and processed. In 2019 and 2022 a global waste management score was published to rank nations based on how well (or not) they are processing their waste. You are being provided with the 2022 dataset based on: https://sensoneo.com/global-waste-index/ Your task is to propose a model that is able to predict the score of a nation based on the available predictors.

For this task you are required to:

1. Create a Jupyter notebook and call it task_01.ipynb
2. Create a Data acquisition section in markdown followed by cells that:
    a. Load the data
    b. Explore the data structure
    c. Preview the first and last few rows
    d. Create a markdown cell in which you identify the predictor variables and the target variable (kindly focus only on numerical variables, you can ignore the Country variable).
3. Create a Data exploration section in markdown followed by cells that:
    a. Display a statistical overview of the numerical variables
    b. Display frequency plots for each of the predictor variables
    c. Display box plots for each of the predictor variables

   d. In a markdown cell provide an interpretation of the frequency plots noting any skewness in the data, then an interpretation of the box plots noting the presence of any outliers.
4. Create an Analysis section in markdown followed by cells that:
   a. Display the correlation across all numerical variables
   b. Undertake regression analysis (simple and multiple) where for each regression model you display a summary of the model (no need to display the model as a line graph).
   c. In a markdown cell provide an interpretation of the correlation between each variable and the target variable only (not of predictor variables across each other).
   d. Provide a recommendation for the ideal model with justification. For the ideal model only, you are to provide an interpretation of:
      i. The p-value and what it means regarding the Null Hypothesis
      ii. The R-squared or Adjusted R-squared value and what it means iii. The F-Statistic in comparison with that of the other models.

Task 02: Classification

We shall be focusing on the AI4I 2020 dataset which is a synthetic dataset about machine failures. The dataset is available at https://doi.org/10.24432/C5HS5C. You are also being provided with two academic papers:

1. **Rayarao and Rayarao 2024** provide a detailed analysis of the dataset and shed some insights on the challenges being faced as well as some inspiration on what visualisations are possible.
2. **Shah et al 2024** document their experimentation on the said dataset using random forests and support vector machines. They undertake a 2-phase analysis, first attempting to classify the machine failure target variable.

For this assignment we shall focus on the binary classification problem of machine failure only. Following are the steps you need to undertake:

1. Create a Jupyter notebook and name it task_02.ipynb
2. Create a markdown cell and set a heading 1 section titled: “AI4I 2020 classification”. You can add some documentation such as the link to the dataset and to the papers.
3. Create another markdown cell with a heading 2 section titled: “Data acquisition”. Add cells to: a. Load the dataset
   b. Preview the structure
   c. Preview the first and last rows.
   d. In a markdown cell outline the license of the dataset and the ethical considerations in using this dataset.
4. Create another markdown cell with a heading 2 titled: “Data exploration”. Add cells to:
   a. Display statistical information about the variables
   b. Display frequency plots for the predictor variables
   c. Display box plots for the predictor variables stratified by the target variable (Machine Failure).
   d. Display the distribution as a percentage of the target variable values as a percentage of the dataset.
5. Create another markdown cell with a heading 2 titled: “Data preprocessing”. Add cells to: a. Remove irrelevant variables
   b. Create a copy of the dataset that retains only the predictors that you will use for classification and the binary target variable Machine failure.
6. Create another markdown cell with a heading 2 titled: “Model creation”. Add cells to:
   a. Split the dataset in the same fashion that Shah et al 2024 did in their first phase of their research. Ensure that both the training and test dataset retain the same distribution of the target variable.

   b. Next you are being requested to focus on two classifiers. One classifier must be the same as what Shah et al 2024 attempted (Random Forest or Support Vector Machine), the other classifier can be anything you want (Logistic Regression, Decision Tree, Multi-layer Perceptron Neural Network, xgBoost, etc.)
   c. Research what hyper parameter tuning is and for the two classifiers you selected, identify the parameters you want to explore. For example, in the case of a Random Forest you want to experiment with different number of estimators, different values of max depth, etc.
   d. Undertake repeated cross validation on the training dataset. Then for the two estimator types (Random Forest and other selected estimator) display the best parameters and best training score.
   e. Use the best models of the two estimators on the test dataset and for each display the confusion matrix and classification report.

7. Create a markdown cell with a heading 2 section titled “Evaluation and recommendation”. In this section you are to:
   a. Criticise the research by Shah et al 2024, focusing on their experiment and findings in first phase.
   b. Based on your findings and the methodology presented here, criticise your own work and findings.
8. Create a markdown cell with a heading 3 section titled “Model exportation”. In this section you are to create a code cell that exports the best model to a pickle file so that it can be utilised in a production system.

N.B. The aim of this research is not to get better results but to criticise your own work. You are expected to reflect on the data splitting strategy, use of repeated cross validation, differences

Grading Criteria

Task 01

	Requirements	**Mark**
AA3.3	Complete steps 1-3a	7
AA4.1	Complete steps 3b-4a	7
KU1.3	Complete step 4b	5
AA3.1	Complete step 4c	7
KU4.2	Complete step 4d	5
Total		31

Task 02

	Requirements	**Mark**
AA2.1	Complete steps 1-5	7
AA2.4	Complete step 6a	7
SE3.4	Complete steps 6b, 6c, 6d and 8	10
KU2.5	Complete step 6e	5
SE4.3	Complete step 7a	10
SE4.4	Complete step 7b	10
Total		49