MCAST

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	03/01/2024	**Deadline Date**	19/01/2024		
Student Name		**ID Number**		**Class / Group**	

Assessment Criteria	Maximum Mark
AA2.1: Analyse a given data set and identify data cleaning issues.	7
AA2.3: Arrange data in preparation for use within a classifier algorithm.	7
KU2.4: 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	68

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 is a home assignment and should be completed in **2 weeks**.
- This assignment carries a total of 68% 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, 1 PowerPoint presentation PPTX and 1 Python Jupyter Notebook.
- You will be requested to present your work to your lecturer during an interview.
- Copying is strictly prohibited, and any students caught will be subject to the respective MCAST Disciplinary Procedures.

Task

You are to perform research in the detection of anomalies from the Shuttle dataset for which a customized version of the dataset and a peer reviewed paper are being provided. The customization made to the dataset does not relate to the structure and general data distribution. The original dataset was amended slightly to permit tasks covered within this assignment; thus you are to use the provided dataset only. A more elaborate explanation of the original dataset is found on the UCI Machine Learning Repository[1]. A peer reviewed academic paper is being provided for reference and comparison purposes[2]. Create a Python Jupyter Notebook that undertakes machine learning research in the identification of a model that classifies the different types of the target variable. The aim is not to achieve 100% accuracy but to follow a proper research methodology. Use the following tasks and grading criteria to guide you through the expected deliverables.

1. **Understand the domain** by visiting the dataset page and reading the paper to research the subject matter.
2. **Data exploration**:
    a. Identify the predictor and target variables.
    b. Determine the number of observations, data types, presence of null values, range of values, and duplicate rows.
3. **Data cleaning and transformation**:
    a. Determine if missing values should be removed or replaced by the respective mean/median.
    b. Determine if any duplicate rows need to be removed.
    c. Determine if any variable is redundant and thus remove.
    d. Convert all variables to an appropriate data type based on their use.
    e. Determine if each variable needs to be normalized in preparation for your classifiers. If that is the case, then normalize your data.
4. **Visualize the data**:
    a. Determine the distribution of the target variable per class.
    b. Plot the box plot of each predictor variable stratified per target variable.
5. **Analyze the data** by investigating the correlation between each predictor and the target variable.
6. **Prepare your dataset** for modelling:

[1] https://archive.ics.uci.edu/dataset/148/statlog+shuttle
[2] https://doi.org/10.14419/ijet.v7i4.5.20079

a. Split the dataset in a training and test datasets. Note that the intention is that all model training is undertaken on the training dataset. The best model is evaluated on the test (holdout) dataset.
b. Ensure that each dataset retains the same ratio of target variable classes.

7. **Train models** that are ideal for this analysis. Each model is to have several parameters fine-tuned, such that if you are using Random Forests, you consider different number of estimators. Please note that 1 of the models created can be covered from the publishing paper, yet the other models need to be completely different models from the published paper (meaning something other than DT and Naïve Bayes).
    a. Undertake K-Fold cross validation for each of the 3 models with several parameters that are fine-tuned.
    b. Illustrate the performance of the best configuration for each model. This should, at the very least, include confusion matrixes with values for Precision, Recall and F1-Scores.
8. **Evaluate models** by providing an interpretation of the results. Recommend the ideal model based on the overall model F1-score, and for the ideal model identify the best fine-tuned parameters.
9. **Test the ideal model** on the holdout dataset:
    a. Make sure that any data cleaning undertaken on the train and validation dataset is also applied to the test dataset to ensure that the datasets have the same structure.
    b. Run the best model with the ideal configuration on the test dataset.
    c. Generate the confusion matrix and ROC curve.
    d. Interpret results.
10. **Present your research** by:
    a. Create a presentation.
    b. Structure your presentation to include a section for each step from steps 1-9 listed here above.
    c. Include a Title and End/Thank You slide.
    d. Deliver the presentation to your lecturer.

Grading Criteria

	Requirements	Mark
AA2.1	Identify nulls	_ / 1
	Remove Nulls	_ / 2
	Remove duplicates	_ / 2
	Scale data accordingly	_ / 2
AA4.1	Illustrate the distribution of target variable per class in a histogram	_ / 2
	Illustrate box plots per predictor, stratified per target class	_ / 5
AA3.3	Illustrate the correlation across each variable	_ / 4
	Identify positively correlated predictors to the target.	_ / 1
	Identify negatively correlated predictors to the target.	_ / 1
	Identify non correlated predictors to the target.	_ / 1
AA2.3	Split data into train, and test	_ / 3
	Use an ideal split ratio for each	_ / 2
	Ensure that each dataset maintains the same target variable class ratios.	_ / 2
SE3.4	Implement 3 classifiers	_ / 6
	Establish which parameters for each classifier can be fine-tuned	_ / 4
SE4.3	Train different configurations of each model (facilitated if using cross validation with hyperparameters)	_ / 5
	Provide the performance of each model configuration and identify best model and best configuration (at least statistical information yet graph illustration are encouraged)	_ / 5
KU4.2	Justify the ideal classifier used based on a confusion matrix interpretation and comparisons.	_ / 5
KU2.4	Run the ideal model on the test (holdout) dataset	_ / 1
	Generate the confusion matrix	_ / 1
	Generate ROC	_ / 1
	Provide interpretation	_ / 2
SE4.4	Prepare a presentation as specified in steps.	_ / 4
	Present your research to your lecturer.	_ / 6
Total		_ / 68