

## GROUP ASSIGNMENT

### DATA ANALYTICS IN CYBER SECURITY

#### Overview of Assignment

The objective of the Group Assignment is to design a Machine Learning (ML) model for detecting network intrusions. The ML-based predictive model will differentiate between normal 'good' connections and various 'bad' connections, commonly referred to as intrusions or attacks. For this assignment, the model will be trained and tested using a modified version of the KDD99 dataset.

The attack types are categorised into four major groups:

1. DoS (Denial-of-Service): Disruption of service availability, e.g., SYN flood.
2. Probe: Surveillance and probing activities, e.g., port scanning.
3. R2L (Remote-to-Local): Unauthorised access from a remote machine, e.g., password guessing.
4. U2R (User-to-Root): Unauthorised escalation to superuser (root) privileges, e.g., buffer overflow attacks.

#### Instructions

1. You will work with the provided **SKML-complete.ipynb** notebook, which utilises the **Boosted NSL** train set and the pre-processed **ppNSL** test set.
2. Individual Task (70% weightage):
  - (a) Each group member must select a distinct classifier category from those suggested in the notebook code, ensuring no two members choose the same category.
  - (b) Create an initial baseline model using your chosen classifier with its default hyperparameters.
  - (c) Create an optimised model from the baseline using one or more of the following optimisation strategies:
    - Hyperparameter tuning with cross-validation.
    - Feature selection via correlation analysis.
    - Feature selection based on feature importance.
    - Addressing class imbalance through resampling or weighting.

- (d) Analyse the significant differences between your baseline model and the optimised model.
- (e) Clearly explain your optimisation process and evaluate its impact on overall model performance

3. Group Task (30% weightage):

- (a) Coordinate within the group to apply consistent scaling and normalisation strategies, ensuring uniformity across all train and test datasets.
- (b) Collaboratively select a set of performance metrics to be used for comparing the optimised models across group members.
- (c) Conduct a comparative analysis of the models' performance and discuss the overall findings in detail.

**Deliverables**

Each group must submit one (1) report only in PDF format (maximum 30 pages, excluding appendices) that includes the following chapters:

- (a) Combined Review: A collective overview of the classifiers selected by each group member, detailing the general category or class of algorithms they belong to (e.g., linear, non-linear, ensemble).
- (b) Integrated Performance Discussion: A comprehensive discussion on the performance of the optimised models, highlighting any notable insights gained from comparing the different models.
- (c) Individual Chapters: Each group member must contribute an individual chapter containing:
  - i. An explanation of the optimisation strategies applied.
  - ii. Evaluation of the significant differences between the baseline and optimised models.
- (d) Each group member must also submit a single JupyterLab notebook as an appendix, containing:
  - i. Classification performance statistics, both pre- and post-optimisation.
  - ii. The implementation of the selected optimisation strategy.
  - iii. Relevant visualisations to support the analysis.

**Report Requirements**

1. Any graphs and charts that appear in the JupyterLab notebook **MUST NOT** appear in the report.
2. No code blocks in the report.
3. Page format: 12-point font, line spacing 1.5, page number at the bottom.
4. Every report must have a front cover with the following details:
  - (a) Names of group members
  - (b) Intake code
  - (c) Module Title
  - (d) Assignment Title
5. Major sections should be consecutively numbered, as well as figures and tables.
6. All information, figures and diagrams obtained from external sources must be referenced using the APA referencing system.
7. Plagiarism is a serious offence and will automatically be awarded zero (0) marks.
8. Reports that are unprofessional in their presentation (disorganized, inconsistent look) will not fare well when marks are allocated.

**Marking Guide****Report: Group Chapters – CLO2****Report organisation and presentation meets requirements**

- 7.5-10 Report format meets all requirements, only minor omissions/errors
- 6.5-7.4 Report format does not meet some requirements
- 5.0-6.4 CAP for copy/paste graphs/charts (not referenced to ipynb submission)
- 2.0-4.9 Serious problems with organisation and presentation

**Overview of the selected algorithms**

- 7.5-10 Drawn from different classes, good review of their characteristics
- 6.5-7.4 Drawn from different classes, acceptable review
- 5.0-6.4 Not drawn from different classes, minimal review, lacks depth
- 2.0-4.9 Missing or superficial overview, serious conceptual problems

Integrated Discussion of the performance of the different models

7.5-10 Excellent discussion, with significant insights and recommendations

6.5-7.4 Good comparison of results, some good insights and recommendations

5.0-6.4 No depth of discussion, simply repeats information from the individual sections.

2.0-4.9 Missing or superficial discussion, serious conceptual problems

**Report: Individual Chapters - CLO3**Explanation of the optimisation strategy applied

10-13 Excellent explanation and justification of parameters chosen

8.5-9.9 Adequate explanation and justification of parameters chosen

6.5-8.4 Minimal explanation and justification of parameters chosen, conceptual problems

0-6.4 Missing or superficial discussion, serious conceptual problems

Analysis of the significance of the difference between the baseline and optimised models

10-13 Excellent discussion, with significant insights and recommendations

8.5-9.9 Good comparison of results, some good insights and recommendations

6.5-8.4 No depth of discussion, simply repeats information from the individual notebook.

0-6.4 Missing or superficial discussion, serious conceptual problems

**JupyterLab Notebook – CLO3**Classification performance (before and after)

14-18 Properly implemented, appropriate metrics, shows creativity/originality

12-13 Properly implemented, appropriate metrics, strictly based on examples

9-11 Indiscriminate copy/paste from examples, conceptual misunderstandings

0-8 Missing or improperly implemented, serious conceptual problems

Model optimisation

10-13 Properly implemented, appropriate parameters, shows creativity/originality

8.5-9.9 Properly implemented, appropriate parameters, strictly based on examples

6.5-8.4 Indiscriminate copy/paste from examples, conceptual misunderstandings

0-6.4 Missing or improperly implemented, serious conceptual problems

Appropriate visualisations

10-13 Properly implemented, uses creative or advanced techniques

8.5-9.9 Properly implemented, strictly based on examples

6.5-8.4 Indiscriminate copy/paste from examples, conceptual misunderstandings

0-6.4 Missing or improperly implemented, serious conceptual problems