

## Assessment Task 4:

This document supplies detailed information on Assessment Task 4 for this unit.

### Key information

- Due: **Sunday 02 October 2022 by 8.00 pm**

### Learning Outcomes

This assessment assesses the following Unit Learning Outcomes (ULO) and related Graduate Learning Outcomes (GLO):

Unit Learning Outcome (ULO)	Graduate Learning Outcome (GLO)
<b>ULO6</b> - Perform model selection and compute relevant evaluation measure for a given problem. <b>ULO7</b> - Use concepts of machine learning algorithms to design solution and compare multiple solutions.	<b>GLO1</b> - Discipline-specific knowledge and capabilities <b>GLO2</b> - Communication <b>GLO4</b> - Critical thinking <b>GLO5</b> - Problem solving <b>GLO6</b> - Self-management

### Purpose

This assessment is an extensive machine learning project. The task is open in nature, where students should make all design decisions to solve a problem and justify their decisions. In addition, they have to design and develop solutions that are better than any existing solutions.

## Assessment 4

### Submission Instructions

- Submit your solution codes into a **notebook file with “.ipynb”** extension. Write discussions and explanations including outputs and figures into a separate file and **submit as a PDF file**.
- Submission other than the above-mentioned file formats will not be assessed and given **zero** for the entire submission.
- Insert your Python code responses into the cell of your submitted “.ipynb” file **followed by the question** i.e., copy the question by adding a cell before the solution cell. If you need multiple cells for better presentation of the code, add question only before the first solution cell.
- Your submitted code should be executable. If your **code does not generate** the submitted solution, then you will **get zero** for that part of the marks.
- Answers must be **relevant and precise**.
- No **hard coding** is allowed. Avoid using specific value that can be calculated from the data provided.
- Use all the **topics covered in the unit** for answering this assignment.
- Submit your assignment **after running each cell individually**.

## Assessment Task 4:

### Questions

---

## Background

In this project you are given a dataset and an article that uses this dataset. The authors have developed eight ML models for cyber security intrusion detection and compared their performance. You must read the article to understand the problem, the dataset, and the methodology to complete the following tasks.

## Dataset

NSL-KDD dataset has been developed to solve problems in KDD 99 challenge. It does not contain unnecessary and repetitive records according to the original KDD 99 data set. A detailed description of the dataset can be found in the Dataset section of the provided article. You can also use other sources for better understanding the dataset and answer questions.

Please use the provided dataset “Intrusion\_detection\_NSL\_KDD.csv” for answering the questions and DO NOT DOWNLOAD AND USE dataset from any other sources. Use the file “FieldNames.pdf” for pre-processing the independent and target variables BEFORE ANSWERING any questions.

## Tasks:

1. Read the article and reproduce the results (Accuracy, Precision, Recall, F-Measure) for NSL-KDD dataset using following classification methods:

- SVM Linear
- SVM Quadratic
- SVM Cubic
- KNN Fine
- KNN Medium
- KNN Cubic
- TREE Fine
- TREE Medium

These results can be found in Table 4 of the manuscript and should be used for comparison purposes, if required. Write a report summarising the dataset, used ML methods, experiment protocol and results including variations, if any. During reproducing the results:

- i) you should use the same set of features used by the authors.
- ii) you should use the same classifier with exact parameter values.
- iii) you should use the same training/test splitting approach as used by the authors.
- iv) you should use the same pre/post processing, if any, used by the authors.

[N.B. Definition of used algorithm can be found in this link: <https://au.mathworks.com/help/stats/choose-a-classifier.html>. However, **your submission must be in python not in Matlab.**]

N.B.

(i) If you find any issue in reproducing results due to incomplete description of model in the provided article then make your own assumption and explain the reason. If your justification is correct then your solution will be considered correct and assessed accordingly.

(ii) If you find some subtle variations in results due to implementation differences of methods used in the study i.e., packages and modules in Python vs Matlab implementation, then appropriate explanation of them will be considered during evaluation of your submission.

## Assessment Task 4: .

(iii) Similarly, variation in results due to randomness of data splitting will also be considered during evaluation based on your explanation.

(iii) Obtained marks will be proportional to the number of ML methods that you will report in your submission with correctly reproduced results.

(iv) Make sure your submitted Python code segment generates the reported results, otherwise you will receive zero marks for this task.

Marking criteria:

- i) Unsatisfactory ( $x < 4$ ): tried to implement the methods but unable to follow the approach presented in the article. Variation of marks in this group will depend on the quality of report.
- ii) Fair ( $4 \leq x < 5$ ): appropriately implemented 50% of the methods presented in the article. Variation of marks in this group will depend on the quality of report.
- iii) Good ( $5 \leq x < 7$ ): appropriately implemented 70% of the methods presented in the article. Variation of marks in this group will depend on the quality of report.
- iv) Excellent ( $x \geq 7$ ): appropriately implemented  $\geq 90\%$  of the methods presented in the article. Variation of marks in this group will depend on the quality of report.

2. Design and develop your own ML solution for this problem. The proposed solution should be different from all approaches mentioned in the provided article. This does not mean that you must have to choose a new ML algorithm. You can develop a novel solution by changing the feature selection approach or parameter optimisations process of used ML methods or using different ML methods or different combinations of them. This means, the proposed system should be substantially different from the methods presented in the article but not limited to only change of ML methods. Compare the result with reported methods in the article. Write a technical report summarising your solution design and outcomes. The report should include:

- i) Motivation behind the proposed solution.
- ii) How the proposed solution is different from existing ones.
- iii) Detail description of the model including all parameters so that any reader can implement your model.
- iv) Description of experimental protocol.
- v) Evaluation metrics.
- vi) Present results using tables and graphs.
- vii) Compare and discuss results with respect to existing literatures.
- viii) Appropriate references ([IEEE numbered](#)).

*N.B. This is a HD (High Distinction) level question. Those students who target HD grade should answer this question (including answering all the above questions). For others, this question is an option. This question aims to demonstrate your expertise in the subject area and the ability to do your own research in the related area.*

Marking criteria:

<b>Quality of solution</b>	<b>Quality of report</b>	<b>Overall score</b>
Unsatisfactory	Unsatisfactory	Unsatisfactory; Score < 5
Unsatisfactory	Fair	Unsatisfactory; Score < 7
Unsatisfactory	Good	Unsatisfactory; Score < 10
Fair	Unsatisfactory	Unsatisfactory; Score < 10
Fair	Fair	Fair; Score < 12
Fair	Good	Fair; Score < 14
Good	Unsatisfactory	Fair; Score < 14

#### Assessment Task 4:

Good	Fair	Good; Score <16
Good	Good	Good; Score >=16

#### Quality of solution

- *Unsatisfactory: an appropriate solution presented whose performance is lower than the reported lowest performance in the article.*
- *Fair: an appropriate solution presented whose performance is at least better than the lowest performance reported in the article.*
- *Good: an appropriate solution presented whose performance is better than the best reported performances in the article.*

#### Quality of report

- *Unsatisfactory: either the report does not include all criteria mentioned above or the quality of description is poor.*
- *Fair: the report has included all criteria mentioned above with an average quality of description.*
- *Good: the report can be considered as a first draft for a publication.*

#### 3. Present your result in a 3 minutes video using PowerPoint slides/animation. (5 marks)

*Marking criteria:*

- Quality of audio presentation*
- Quality of slides/animation.*
- Completeness of the information.*

