BAHRAIN POLYTECHNIC
بـولـيـتـكـنـك البحريـــن

Assessment Cover Sheet

Assessment Title	Project _Group		
Assessment Type	Uncontrolled	Group	Not must-pass
Due Date	15-12-2025	**Course Code**	IT7009
Course Title	Artificial Intelligence		
Internal Moderator's Name	Mahmoud Alhamad		
External Examiner's Name			

Instructions:

1. This cover sheet must be completed (section in blue below) and attached to your assessment before submission in soft copy.
2. The deadline for this assessment is 15-Dec-2025 (11:55 PM).
3. This assessment carries 35% marks distributed to a total of 3 questions assessing CILO 1, CILO2, and CILO 3.
4. Late submission, marks will be capped at 60%. Two calendar days after the deadline, students will get 0%.
5. References consulted (if any) must be properly acknowledged and cited.
6. The assessment has a total of 5 pages.

Learner ID		**Date Submitted**	
Learner Name			
Programme Code			
Programme Title			
Lecturer's Name			

By submitting this assessment for marking, I affirm that this assessment is my own work.

Do not write beyond this line. For assessor use only.

Learner Signature

Assessor's Name			
Marking Date		**Maks Obtained**	

Comments:

Submissions Instructions

- Only the team-leader of the group must submit the project deliverables.
- For the given problem statement of project, student needs to build project on Google Collaboratory and upload the downloaded. ipynb file on the submission link which will be available on the Moodle.
- Project submission must be supported with the report where the students must explain the steps required for the completion of the project as per the details given under the "Project Report" section at the end of the project.
- A project report must be in the ***document*** format (doc. or docx). Groups are allowed to make submission only as one Zipped folder containing. ipynb files, and report in docx format.
- Project Zip folder for submission must be named with stream number with only group leader student ID. If stream number is 08 and group leader student ID is 202316152, the zipped folder must be named as "08 _202316152".
- Any submission with ***RAR*** format is not allowed and the student will be awarded ***zero*** mark for that submission. Please include a cover page within the report with the full name and student ID of all group members.
- In Google Collaboratory, each of the steps performed for the project must be supported with the comments or explanations.
- The submission needs to be uploaded on the course **Moodle page**.

Project Scenario:

Modern enterprises face an overwhelming volume of network activity, making manual threat detection increasingly unsustainable. To enhance response time and reduce analyst fatigue, cybersecurity teams are turning to machine learning to automate the classification of network flows and identify potential threats.

This project simulates the development of an intelligent threat detection pipeline, designed to support real-world cybersecurity operations. The goal is to build a machine learning powered solution that analyzes structured network data and identifies potential threats which results in enhancing response efficiency, prioritizing analyst efforts, and reducing time-to-detection for high-risk behaviors.

Dataset Overview:

The dataset provided consists of labeled instances of network flow sessions, each described by structured numeric features:

- **Numeric Data:**
  - **Network identifiers:** Source/Destination IP and Port, protocol type, timestamp.
  - **Traffic behavior:** Packet counts (forward/backward), bytes transferred, flow rates.
  - **Temporal indicators:** Duration, inter-arrival times, active/idle durations.

    - **Protocol-level signals:** TCP flag counts and similar indicators.

- **Target Label:**

  The dataset contains pre-labeled samples, with each flow session assigned threat categories based on observed behavior in the network. These categories represent a range of activities, from typical benign traffic to patterns commonly associated with suspicious or malicious intent.
  Your model should be trained using the provided label structure. Optionally, you may choose to re-group or simplify the labels (e.g., merge into broader categories) if it improves performance or interpretability.

Project Objectives:

The project centers on the design and evaluation of a machine learning pipeline for threat detection using structured network flow data. The primary objectives are as follows:

1. **Supervised Threat Classification:**

   Design and implement supervised machine learning models to categorize network flow records into their corresponding threat levels. The approach should include a well-reasoned selection of algorithms, careful construction of input features, and a structured training process. Performance should be assessed using appropriate classification metrics that are discussed in the course, with results critically analyzed in terms of their reliability, trade-offs, and relevance to cybersecurity contexts.

2. **Anomaly Detection using Unsupervised Learning:**

   Explore unsupervised learning approaches to uncover network flow sessions that deviate from established behavioral patterns. This involves selecting appropriate techniques for identifying outliers, establishing meaningful criteria for anomaly detection, and interpreting the outcomes with consideration for their potential significance in a cybersecurity context.

3. **Feature Analysis and Interpretability:**

   Examine the influence of input features on model predictions and assess their significance in distinguishing between threat categories. Highlight the importance of interpretability in cybersecurity, where understanding model behavior is essential for analyst, informed decision-making, and actionable insights.

4. **Critical Evaluation and Reflection:**

Evaluate the overall approach in terms of effectiveness, limitations, and practical applicability. Reflect on modeling decisions, data challenges, and areas for improvement. Compare the outcomes of supervised and unsupervised methods and provide a reasoned summary of insights and future recommendations.

Suggested Technical Directions:

The project should demonstrate a well-reasoned approach to data preparation, modeling, and evaluation. Students are expected to clearly explain and justify their methodological choices at each stage of the pipeline:

1. **Data Preparation:**

   Prepare the dataset by addressing issues such as data quality, consistency, and formatting. Preprocessing decisions should be explained with reference to their impact on model readiness and alignment with the intended learning approach.

2. **Feature Development:**

   Identify and construct relevant features that support effective model learning. This may include transforming, selecting, or organizing features in a way that enhances model performance or interpretability. All transformations must be documented and justified.

3. **Model Design:**

   Implement and evaluate appropriate supervised and unsupervised learning models based on project objectives. The choice of models should be supported with reasoning related to the dataset structure, task type, and expected outputs. Approaches to model tuning, input handling, and architecture should be effectively outlined.

4. **Evaluation & Validation:**

   Evaluation should involve the use of well-suited metrics that reflect the specific goals of the project, such as accuracy, precision, recall, or others as appropriate. Different modeling approaches or techniques should be compared effectively, with attention to their strengths, weaknesses, and practical implications. In cases involving unsupervised learning, the basis for identifying meaningful patterns or outliers should be clearly explained and supported with reasoning. Overall, the evaluation should include critical analysis of the results, acknowledge any limitations and reflect on how the chosen methods align with the problem being addressed.

Project Deliverables:

1. **Codebase:**
   Modular and well-documented implementation provided in **Jupyter Notebook format (ipynb)**. The notebooks should clearly demonstrate the full workflow, including data preprocessing, model training, evaluation, and result visualization. All code should be organized and annotated to ensure clarity and ease of understanding.

2. **Project Report:**

   The report should be well structured and professionally formatted, addressing the following aspects:

   a. A clear explanation of the problem and its relevance within a real-world or domain-specific context.
   b. Statement of project objectives and the AI-based approach adopted to address them.
   c. Description of the data used, and steps taken to prepare it for analysis or modeling.
   d. Explanation of the machine learning methods applied, including rationale for choices made.
   e. Evaluation of model performance results, with comparison across approaches.
   f. Interpretation of model behavior, including analysis of influential features or decision-making patterns.
   g. Discussion of practical implications, challenges encountered, limitations of the approach, and possible improvements.
   h. Summary of overall findings and final reflections.