Team Members: Florisita Mali, Gresja Kulejmani, Ledio Hima

Topic: Multi-Modal Phishing Detector Agent (Websites and Emails)

Introduction:

The problem addressed in this project is to design an intelligent agent that can perceive website-related features, assess the likelihood that a website is phishing, and take an appropriate protective action on behalf of the user. The agent must operate under partial observability, make decisions based on uncertain information, and minimize harmful outcomes, particularly allowing phishing websites to be accessed. Rather than only producing a classification label, the agent is required to select rational action - blocking, warning, or allowing access - based on its assessment of risk.

In addition to website phishing detection, the agent is extended to analyze email-based phishing attacks, which represent one of the most common phishing vectors.
Key characteristics: This dataset contains 4 features extracted from 5000 phishing webpages and 5000 legitimate webpages,

Dataset Description: *This project uses two separate datasets for both Email and Website Phishing detection. Details about both datasets are provided associated with this report.*

INTELLIGENT AGENT FORMULATION:

***Agent Type*:** The proposed system is a *Learning Agent with Logical Reasoning*, combining: a learned predictive model, a rule-based decision mechanism, for both websites and emails using the same reasoning logic.

The agent performance is measured how correctly it classifies the websites and emails as Phishing or not. Also for each prediction there will be shown: the probability score, risk level and confidence visualization. On another section there will be show the proposed action like:

- ALLOW ACCESS - Phishing probability is below the defined risk threshold
- WARNING - Phishing probability falls within a medium-risk range
- BLOCK ACCESS - Phishing probability exceeds a high-risk threshold

Standard classification metrics are going to be calculated on a test dataset: *Accuracy, Precision, Recall, F1-score (primary metric).*

The agent operates in a *partially observable, static and discrete environment.*

- Partially observable because the agent only has access to extracted website features, rather than the full intent or strategy of the attacker. For emails: only extracted content, metadata, and structural features are available
- Static since each website or email is analyzed independently and does not change during the decision-making process
- Discrete because both the input features and the possible actions of the agent are represented using discrete values.

How it works:

When a user enters a website URL or an email message, the intelligent agent perceives it as raw environmental input and automatically extracts relevant lexical and structural features. There will be use the word embendings These features are preprocessed and passed to a trained machine learning model, which produces a phishing probability representing the agent's belief about the website's legitimacy. This belief is then evaluated using a rule-based reasoning module that determines the appropriate protective action. Based on the reasoning outcome, the agent selects and presents one of three actions—allowing access, warning the user, or blocking the website or email—along with a brief explanation, thereby completing the perception–reasoning–action cycle of an intelligent agent.

AI Techniques to Be Used

To satisfy the requirement of integrating multiple artificial intelligence techniques, this project implements two distinct paradigms taught in the course*: statistical learning (machine learning)* and *logical reasoning (rule-based systems)*.

1. **Statistical Learning (Machine Learning)**

Supervised classifiers are trained to detect phishing in both websites and emails *using Decision Tree* and *Random Forest* models. For websites, the models use structured URL- and content-based features that are directly suitable for tree-based learning. For emails, *TF-IDF vectorization* is applied to transform textual content into numerical features before classification. Decision Trees provide interpretability, while Random Forest is used as the primary model to improve robustness and reduce overfitting.

2. **Logical Reasoning (Rule-Based System)**

The output of the classifier (probability of phishing) will be passed to a rule-based reasoning module that determines the agent's action. For example:

- If phishing probability > 0.80 → Block access
- If phishing probability between 0.50 and 0.80 → Warn user
- If phishing probability < 0.50 → Allow access

This logical reasoning layer ensures rational decision-making and enhances system transparency.

Why these techniques

Decision Tree and Random Forest models were chosen due to their low computational complexity, fast training time on structured phishing data, minimal hyperparameter tuning, and strong interpretability. In contrast, SVMs and neural networks typically require higher computational cost, extensive parameter optimization, and longer training times, while offering limited transparency. The rule-based layer further improves interpretability by converting model outputs into explicit, policy-driven actions, making the overall system efficient, explainable, and suitable for an intelligent anti-phishing agent.