AIAP® Batch 22 Technical Assessment

Deadline: **1900 hrs, 1st December 2025**

Tasks

This assessment consists of two parts:

1. Exploratory Data Analysis in Jupyter Notebook
2. End-to-end Machine Learning Pipeline in Python Scripts (`.py`)

Technical Assessment Overview

There are two parts to the Technical Assessment: Exploratory Data Analysis and End-to-end Machine Learning Pipeline. You are to attempt both parts and submit the deliverables by uploading them to your own **private** GitHub repository. The following flowchart outlines the major steps for the Technical Assessment. Details will be provided in the subsequent sections of this document.
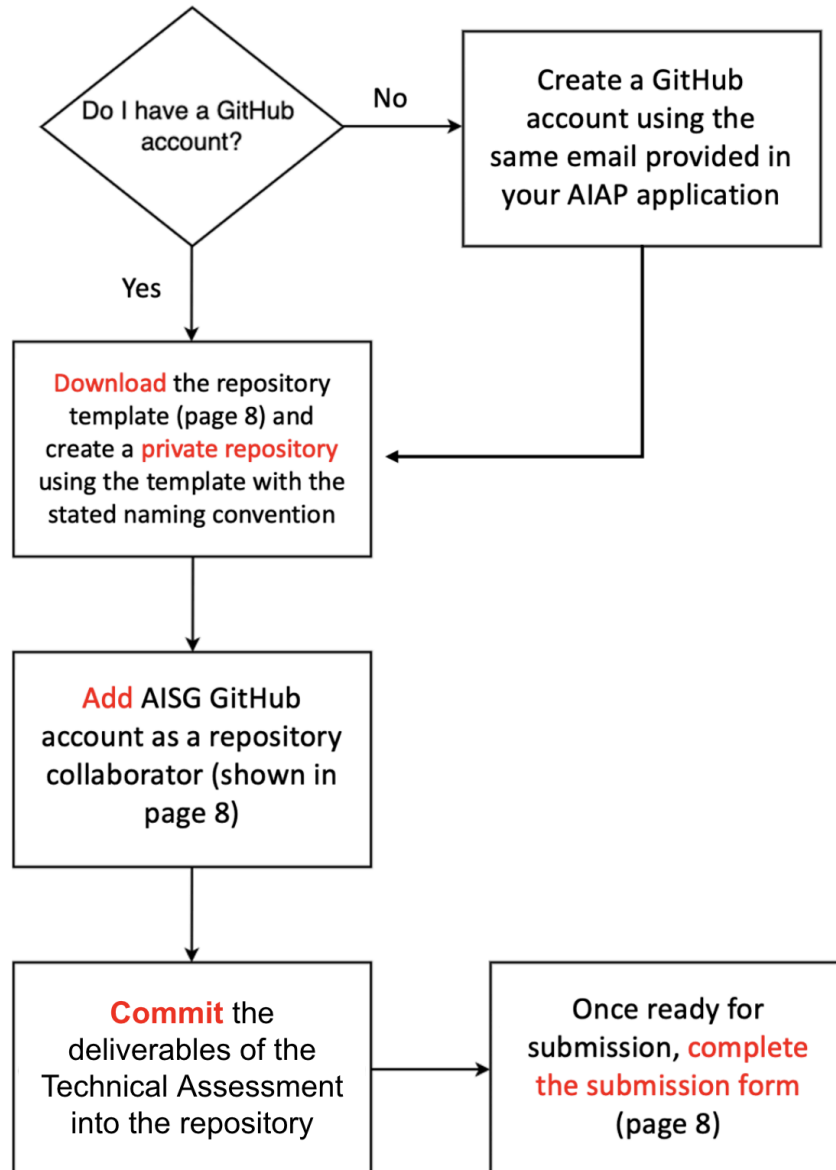
Task 1 - Exploratory Data Analysis (EDA)

Using the dataset specified in the **Dataset** section at page 6, conduct an EDA and create an interactive notebook (.ipynb file) in **Python** that can be used as a presentation to explain the findings of your analysis. It should contain appropriate visualizations and explanations to assist readers in understanding how these elaborations are arrived at and their implications.

Deliverable
1. Jupyter Notebook in **Python**: a `.ipynb` file named **`eda.ipynb`**. (do adhere to the naming requirement)

Evaluation
In the submitted notebook, you are required to

1. Outline the steps taken in the EDA process
2. Explain the purpose of each step
3. Explain the conclusions drawn from each step
4. Explain the interpretation of the various statistics generated and how they impact your analysis
5. Generate clear, meaningful, and understandable visualizations that support your findings
6. Organize the notebook so that is it clear and easy to understand

Please note that your submission will be heavily penalized for any of the following conditions:

1. `.ipynb` missing in the submitted repository
2. `.ipynb` cannot be opened on Jupyter Notebook
3. Explanations missing or unclear in the submitted Jupyter Notebook

Task 2: End-to-end Machine Learning Pipeline

Design and create a machine learning pipeline (MLP) in Python scripts (`.py` files) that will ingest and process the entailed dataset, subsequently, feeding it into the machine learning algorithm(s) of your choice.

Do not develop your MLP in an interactive notebook.

The pipeline should be easily configurable to enable easy experimentation of different algorithms and parameters as well as ways of processing data. You can consider the usage of a config file, environment variables, or command line parameters.

Within the pipeline, data (provided in the Dataset section, Page 6) must be fetched/imported using SQLite, or any similar packages.

Deliverables
1. A folder named `**src**` containing Python modules/classes in `.py` format.
2. An executable bash script `**run.sh**` at the base folder of your submission to run the aforementioned modules/classes/scripts. DO NOT install your dependencies in the `run.sh`; this will be taken care of automatically when we assess the assignment if you have created your `requirements.txt` correctly.
3. A `**requirements.txt**` file in the base folder of your submission.
4. A `**README.md**` file that sufficiently explains the pipeline design and its usage. You are required to explain the thought process behind your submitted pipeline in the README. The README is expected to contain the following:
    a. Full name (as in NRIC) and email address (stated in your application form).
    b. Overview of the submitted folder and the folder structure.
    c. Instructions for executing the pipeline and modifying any parameters.
    d. Description of logical steps/flow of the pipeline. If you find it useful, please feel free to include suitable visualization aids (eg, flow charts) within the README.
    e. Overview of key findings from the EDA conducted in Task 1 and the choices made in the pipeline based on these findings, particularly any feature engineering. Please keep the details of the EDA in the `.ipynb`. The information in the `README.md` should be a quick summary of the details from `.ipynb`.
    f. Describe how the features in the dataset are processed (summarized in a table).
    g. Explanation of your choice of models for each machine learning task.
    h. Evaluation of the models developed. Any metrics used in the evaluation should also be explained.
    i. Other considerations for deploying the models developed.

Evaluation

The submitted MLP, including the `README.md`, will be used to assess your understanding of machine learning models/algorithms as well as your ability to design and develop a machine learning pipeline. Specifically, you will be assessed on

1. Appropriate data preprocessing and feature engineering
2. Appropriate use and optimization of algorithms/models
3. Appropriate explanation for the choice of algorithms/models
4. Appropriate use of evaluation metrics
5. Appropriate explanation for the choice of evaluation metrics
6. Understanding of the different components in the machine learning pipeline

In your submitted Python scripts (`.py` files), you will be assessed on the quality of your code in terms of reusability, readability, and self-explanatory.

Please note that your submission will be penalized for any of the following conditions:

1. Incorrect format for `requirements.txt`
2. `run.sh` fails upon execution
3. Poorly structured `README.md`
4. Disorganized code that fails to make use of functions and/or classes for reusability
5. MLP not submitted in Python scripts (`.py` files), including MLP built using Jupyter Notebooks.

Note for Windows users

DO NOT submit a Windows batch (`*.bat`) script in replacement of the bash script. Use either 'Windows Subsystem for Linux (WSL)' or 'Git Bash'/'cygwin' for the creation of the bash script.

Problem Statement

Objectives

As a new hire to the ML Engineering team at CyberProtect, you have been entrusted with a critical task: protecting the users of CyberProtect from phishing attacks when they are assessing websites on the internet.

CyberProtect has collected a database of phishing websites and legitimate websites. CyberProtect's data engineering team has extracted various features from these websites that may be helpful to your tasks. These features are described below in the "List of Attributes" section.

By leveraging these features, you will develop predictive models that can be installed as an extension and predict if a website is a phishing attack and warn the user before allowing the user to access it.

Specifically, your task is to build and evaluate prediction models, and also identify their respective key features of the dataset that categorise whether the website is a phishing attack. Your analysis should also include evaluation of which features contribute most significantly to phishing attack prediction.

In your submission, you are expected to build and evaluate **at least three suitable models** for this task and justify your choices based on the dataset provided.

Dataset

The dataset contains various features such as number of URL redirects, number of external references , which are gathered by the data engineering team to help your work. Specific information on the various dataset features shall be provided in the next page.

Important Note: The dataset contains synthetic or contaminated data. Therefore, you should **state clearly any assumptions or justifications that you make in processing the data, including handling of outliers, missing values, and data quality issues.**

You can query the dataset using the following URL:
https://techassessment.blob.core.windows.net/aiap22-assessment-data/phishing.db

Instructions for querying the database

The dataset can be accessed through the `phishing.db`. You may find either of the following packages, `SQLite` or `SQLAlchemy`, useful for accessing this database.

You should place the `phishing.db` file in a `data` folder. Your machine learning pipeline should retrieve the dataset using the relative path `data/phishing.db`.

DO NOT upload the `phishing.db` onto your GitHub repository.

List of Attributes

Attribute	Description
LineOfCode	Number of lines of code
LargestLineLength	Longest line of code
NoOfURLRedirect	Number of URL redirect from the website
NoOfSelfRedirect	Number of Self redirect from the website
NoOfPopup	Number of Pop ups from the website
NoOfiFrame	Number of iFrames from the website
NoOfSelfRef	Number of clickable links between the same domain
NoOfExternalRef	Number of clickable links to external websites
Robots	Does the website have a robot.txt
IsResponsive	Is the website can be appropriately adapted across devices
Industry	The industry that the website belongs to
DomainAgeMonths	Number of Months since the domain has been created
HostingProvider	The hosting provider of the website
label	0 means phishing website and 1 means legitimate website

Submission Format

Create a [GitHub](#) account using the **same** email provided in your AIAP application form. Download the repository template from:
https://techassessment.blob.core.windows.net/aiap22-assessment-data/aiap22-NAME-NRIC.zip

The downloaded repository template contains a hidden folder: `.github`. The `.github` folder contains scripts to execute your end-to-end machine learning pipeline using GitHub Actions. Specifically, it will first install the required dependencies using your `requirements.txt` and subsequently, execute your bash script (`run.sh`). You can manually trigger the pipeline under Actions in your repository.

Using the downloaded template, create a **private** repository using the following naming convention:
> **aiap22-<full name (as in NRIC) separated by dashes>-<last 4 characters of NRIC>**

For example, `aiap22-john-lim-der-hui-321A`. Candidates who do not adhere strictly to this naming convention may be penalized for their code submission.

Ensure your intended code submission is in the **main** branch of your remote Github repository.

Add the following account as a collaborator in your private repository:
- Username: **AISG-AIAP**
- Email: **aiap-internal@aisingapore.org**

Your repository is to have the following structure:

```
├────── .github
├────── src
│       └────── (python files constituting the end-to-end ML pipeline in .py format)
├────── README.md
├────── eda.ipynb
├────── requirements.txt
└────── run.sh
```

We encourage you to adhere to Git best practices. Once your repository is ready for submission, complete the following form at https://forms.gle/z2USSMNrxrTHsvwdA

NOTE: During the assessment period, you are still allowed to make changes to your repository if you submit the Google form before the deadline.
After the deadline, do **not** make any further changes to your code repository. Candidates who do not adhere strictly to this instruction can be penalized for their code submission.