AIIP 5 Technical Assessment

Deadline: **1900 hrs, 17th March 2025**

Tasks

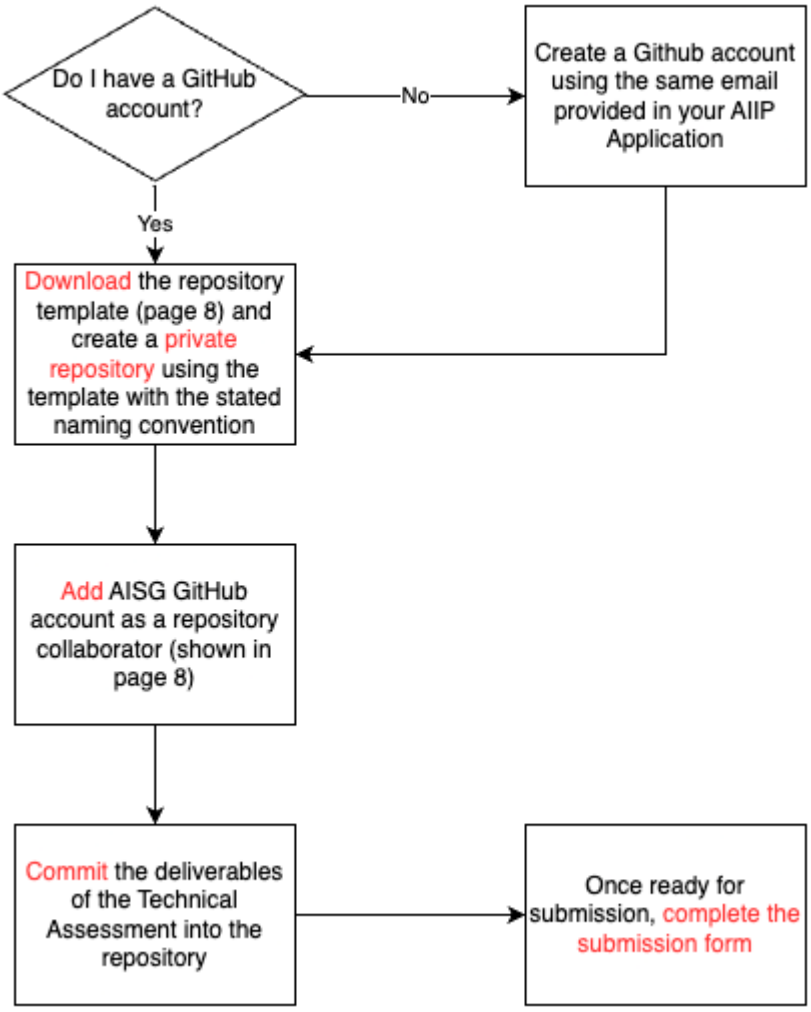
This assessment consists of two parts:

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

AI Singapore
innovation 4.0, 3 Research Link #02-04 Singapore 117602 | *www.aisingapore.org*

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 visualisations 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. Organise the notebook so that is it clear and easy to understand

Please note that your submission will be heavily penalised 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 visualisation 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 (summarised 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 penalised for any of the following conditions:

1. Incorrect format for `requirements.txt`
2. `run.sh` fails upon execution
3. Poorly structured `README.md`
4. Disorganised 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

A leading agri-tech company, AgroTech Innovations, faces significant challenges in optimising crop yields and resource management due to inefficiencies in their controlled environment farming systems. Despite having advanced sensor technologies, the company seeks to enhance its operational efficiency and support future research and development initiatives.

As a machine learning engineer at AgroTech Innovations, you are tasked with developing machine learning models to address these challenges. Your goal is to create models to **predict the temperature conditions** within the farm's closed environment, ensuring optimal plant growth. Additionally, you will develop models to **categorise the combined "Plant Type-Stage"** based on sensor data, aiding in strategic planning and resource allocation.

By implementing these models, you will help AgroTech Innovations improve crop management, optimise resource usage, and increase yield predictability. These efforts will not only enhance current operations but also provide valuable insights for future agricultural innovations.

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

Dataset

The dataset contains sensor readings and plant-related information collected from various agricultural zones within a controlled environment. It includes features such as temperature, humidity, light intensity, CO_2 levels, and nutrient concentrations. Note that the dataset may contain synthetic data. Therefore, you would need to state any assumptions you make.

You can query the datasets using the following URL:
https://techassessment.blob.core.windows.net/aiip5-assessment-data/agri.db

Instructions for setting up SQLite and querying the database

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

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

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

List of Attributes

Attribute	Description
System Location Code	Code representing the specific zone or area within the farm.
Previous Cycle Plant Type	Type of plant grown in the previous cycle in the same location.
Plant Type	Current type of plant being grown.
Plant Stage	Current growth stage of the plant (e.g., seedling, vegetative, maturity).
Temperature Sensor (°C)	Temperature reading from the sensor, measured in degrees Celsius.
Humidity Sensor (%)	Humidity level as measured by the sensor, in percentage.
Light Intensity Sensor (lux)	Light intensity measured by the sensor, in lux.
CO2 Sensor (ppm)	Carbon dioxide concentration measured by the sensor, in parts per million.
EC Sensor (dS/m)	Electrical conductivity measured by the sensor, in decisiemens per meter.
O2 Sensor (ppm)	Oxygen concentration measured by the sensor, in parts per million.
Nutrient N Sensor (ppm)	Nitrogen concentration measured by the sensor, in parts per million.
Nutrient P Sensor (ppm)	Phosphorus concentration measured by the sensor, in parts per million.
Nutrient K Sensor (ppm)	Potassium concentration measured by the sensor, in parts per million.
pH Sensor	pH level measured by the sensor.
Water Level Sensor (mm)	Water level measurement by the sensor, in millimeters.

Submission Format

Create a GitHub account using the **same** email provided in your AIIP application form.

Download the repository template from:
https://techassessment.blob.core.windows.net/aiip-intake5-assessment-data/aiip5-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:

aiip5-<full name (as in NRIC) separated by dashes>-<last 4 characters of NRIC>

For example, `*aiip5-john-lim-der-hui-321A*`

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/oxEXY8SiVN22Z9Ne9

NOTE: During the assessment period, you are still allowed to make changes to your repository after submitting the form.