Technical Assessment

Deadline: **1900hrs, 17th January 2022**

Submissions after the deadline will not be accepted. Please submit the completed assessment early to ensure a smooth submission process.

Tasks

This technical assessment consists of two main parts:

1. Exploratory Data Analysis (EDA)
2. End-to-end Machine Learning Pipeline (MLP)

You are to attempt both parts and package a submission containing deliverables for each of the tasks.

Task 1 - Exploratory Data Analysis (EDA)

Using the dataset specified in the **Data** section below, conduct an EDA and create an interactive notebook 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 as well as their implications.

Deliverable

1. Notebook in **Python**: An `.ipynb` file named **`eda.ipynb`**.

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 visualisations that support your findings
6. Organise the notebook so that is it clear and easy to understand

You will be assessed on the usefulness and clarity of visualisations, accuracy and depth of your insights, presentation flow, and structure of your analysis.

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

1. .ipynb missing from submission folder
2. .ipynb cannot be opened on Jupyter Notebook
3. Explanations missing or unclear in the submitted .ipynb

Task 2: End-to-end Machine Learning Pipeline

Design and create a simple machine learning pipeline that will ingest/process the entailed dataset and feed it into the machine learning algorithm(s) of your choice.

The pipeline should be easily configurable to enable easy experimentation of different algorithms and parameters as well as different ways of processing data (e.g. usage of a config file, environment variables, or command line parameters). Within the pipeline, data must be fetched/imported using SQLite, or any similar packages (provided in the `Data` section).

Deliverables

1. A folder named `src` containing Python modules/classes.
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 at 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.
    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`, this section should be a quick summary.
    f. Explanation of your choice of models for each machine learning task.
    g. Evaluation of the models developed. Any metrics used in the evaluation should also be explained.
    h. Other considerations for deploying the models developed.

Evaluation

The submitted README will be used to assess your understanding of machine learning models / algorithms and ability to design and develop a machine learning pipeline. In particular, you will be assessed on

1. Appropriate use of algorithms/models
2. Appropriate explanation for the choice of algorithms/models
3. Appropriate use of evaluation metrics
4. Appropriate explanation for the choice of evaluation metrics
5. Understanding of the different components in the machine learning pipeline

In your submitted Python scripts, you will be assessed on the quality of your code in terms of organisation, structure, clean separation of functionality and ease of use. Code reusability between the two tasks will be viewed favourably.

Please note that your submission will be heavily 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

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 creation of the bash script.

Data

URL

https://techassessment.blob.core.windows.net/aiap10-assessment-data/survive.db

Instructions for setting up SQLite and querying the database

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

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

DO NOT submit the `**survive.db**` in your final submission.

```
├── src
│   └── (relevant files)
├── data
│   └── survive.db
├── README.md
├── eda.ipynb
├── requirements.txt
└── run.sh
```

Objectives

The objective is to predict the survival of coronary artery disease patients using the dataset provided to help doctors to formulate preemptive medical treatments. In your submission, you are to evaluate at least 3 suitable models for estimating the patients' survivals.

Dataset description

The dataset contains the medical records of coronary artery disease patients for a particular hospital. Do note that there could be synthetic features in the dataset. Hence, please ensure that you state and verify any assumptions that you make.

List of Attributes

Attribute	Description
ID	Unique ID for each patient
Survive	If the patient survives: 0 = No , 1 = Yes
Gender	Gender type
Smoke	If the patient smokes
Diabetes	Diabetes conditions of patient
Age	Age of the patient
Ejection Fraction	Strength of heart
Sodium	Level of sodium in the blood serum (mg/dL)
Creatinine	Level of creatinine in the blood serum (mEq/L)
Platelets	Number of platelets in the blood serum (platelets/mL)
Creatine phosphokinase	Level of the enzyme in the blood (mcg/L)
Blood Pressure	Level of blood pressure (mmHg)
Hemoglobin	Level of hemoglobin in the blood (g/dL)
Height	Height of patient (cm)
Weight	Weight of patient (kg)
Favourite color	Favourite color of patient

Submission Format

Your work should be uploaded as a `*.zip` archive to AI Singapore's designated blob store (detailed below). The archive file is to be provided with the following naming convention:

`<full name (as in NRIC) separated by underscores>_<last 4 characters of NRIC>.zip` e.g. `john_lim_der_hui_321A.zip`.

The submission folder is to have the following structure (**as an example**):

```
├── src
│   └── (relevant files)
├── README.md
├── eda.ipynb
├── requirements.txt
└── run.sh
```

Once you have packaged your submission, you are to upload your submission by following the steps detailed below:

1. Download the Azure **azcopy** tool.

2. Use the URL attached in *submission_url.txt* and the **azcopy** tool to upload your files through the command line. You are expected to follow the instructions under "**Option 2: Use a SAS token**" to make use of the **azcopy** tool. No other steps are required. You do not need to log into an Azure account or obtain a subscription to use the tool. The URL includes the required SAS token. Please ensure that you copy the link correctly and remove any white spaces.

3. If your file has been successfully uploaded, you should observe an output that is similar to what is shown below:

```
Job cfebd42e-c333-9143-56aa-ed28b802d9dd summary
Elapsed Time (Minutes): 0.0334
Total Number Of Transfers: 1
Number of Transfers Completed: 1
Number of Transfers Failed: 0
Number of Transfers Skipped: 0
TotalBytesTransferred: 58651
Final Job Status: Completed
```

Note: The ability to use this tool is considered as part of the technical assessment and evaluation. There will be automated checks that will assess the conformance of your uploaded submission to aforementioned specified instructions. Non-conformance to specified conventions/formats will negatively impact your overall score.