AIAP® Technical Assessment Past Years Series:

Student Score Prediction

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

Objectives

You are part of an AI Engineering team in AI Singapore (AISG) that builds AI solutions to solve problem statements from the public and private sectors. Your current client is U.A Secondary School, a public educational institute in Singapore.

U.A Secondary School wants AISG to build a model that can predict the students' O-level mathematics examination scores to identify weaker students prior to the examination timely. Additional support can then be rendered to the students to ensure they are more prepared for the upcoming test. You are given access to U.A Secondary School's past students' performance dataset (a link to retrieve the dataset can be found in the Data section below).

With the given dataset, you are to perform the following two tasks:

1. Exploratory Data Analysis
2. Build an End-to-end Machine Learning Pipeline

Task 1 - Exploratory Data Analysis (EDA)

Using the given dataset, which can be found 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. You should employ appropriate visualizations and statistical techniques to derive meaningful and relevant insights from the dataset.

Deliverables

1. Notebook in **Python**: An `.ipynb` file named **`eda.ipynb`**.
2. Your EDA should:
    - Outline the steps taken in the EDA process
    - Explain the purpose of each step
    - Explain the conclusions drawn from each step
    - Explain the interpretation of the various statistics generated and how they impact your analysis
    - Generate clear, meaningful, and understandable visualizations that support your findings
    - Organize the notebook so that is it clear and easy to understand

Task 2: End-to-end Machine Learning Pipeline (MLP)

Design and create a machine learning pipeline 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. The MLP should entails:

- Appropriate data preprocessing and feature engineering
- Appropriate use and optimization of algorithms/models (at least 3 models)
- Appropriate explanation for the choice of algorithms/models
- Appropriate use of evaluation metrics
- Appropriate explanation for the choice of evaluation metrics

Do ensure the quality of your submitted Python scripts (`.py` files) in terms of reusability, readability, and self-explanatory.

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 Data 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 used in MLP.
2. A `requirements.txt` file in the base folder of your submission.
3. A `README.md` file that sufficiently explains the pipeline design and its usage. You are required to explain the thought process behind your 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.
h. Evaluation of the models developed. Any metrics used in the evaluation should also be explained.

Data

URL for Querying

https://techassessment.blob.core.windows.net/aiap-preparatory-bootcamp/score.db

Instructions for setting up SQLite and querying the database

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

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

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

List of attributes

Attribute	Description
student_id	Unique ID for each student
number_of_siblings	Number of siblings
direct_admission	Mode of entering the school
CCA	Enrolled CCA
learning_style	Primary learning style
tuition	Indication of whether the student has a tuition
final_test	Student's O-level mathematics examination score
n_male	Number of male classmates
n_female	Number of female classmates
gender	Gender type
age	Age of the student
hours_per_week	Number of hours student studies per week
attendance_rate	Attendance rate of the student (%)
sleep_time	Daily sleeping time (hour:minutes)
wake_time	Daily waking up time (hour:minutes)
mode_of_transport	Mode of transport to school
bag_color	Colour of student's bag