Kaggle tutorial
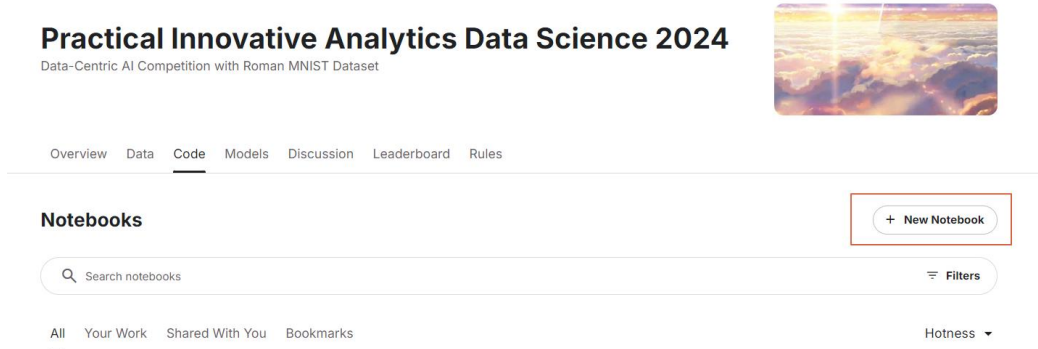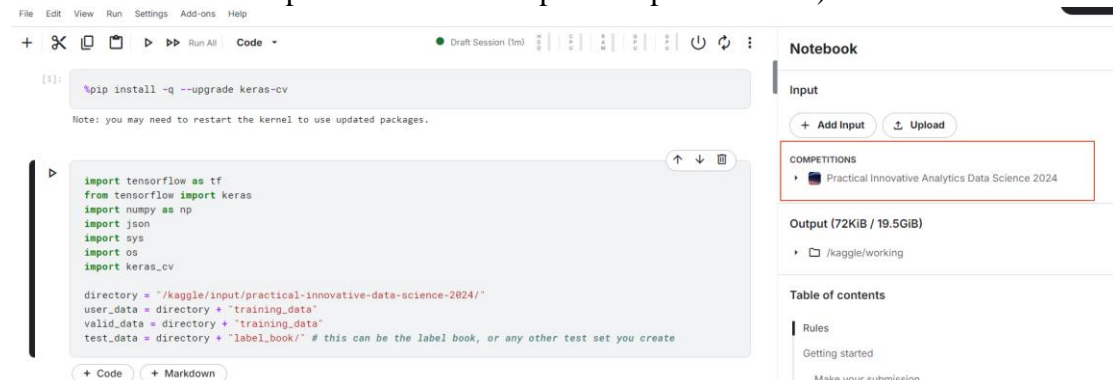
1. **How to get the data?**

   Register [Kaggle](#) first and on the [competition website](#), in the code panel click "New Notebook":

   ## Practical Innovative Analytics Data Science 2024
   Data-Centric AI Competition with Roman MNIST Dataset

   Overview   Data   **Code**   Models   Discussion   Leaderboard   Rules

   Notebooks   + New Notebook

   🔍 Search notebooks   ≡ Filters

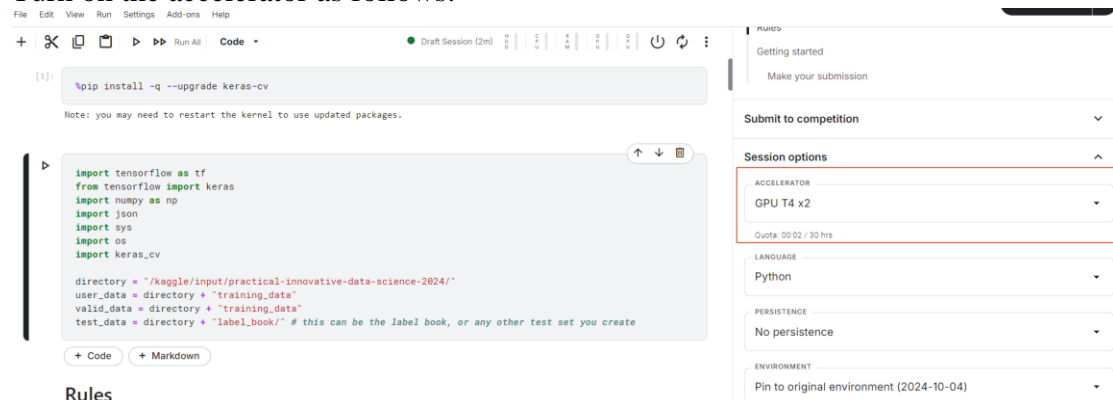   All   Your Work   Shared With You   Bookmarks   Hotness ▾

   The data will be automatically downloaded to the "/kaggle/input" directory (You can also click on file->import notebook to import sample notebook):

   ```
   %pip install -q --upgrade keras-cv

   Note: you may need to restart the kernel to use updated packages.

   import tensorflow as tf
   from tensorflow import keras
   import numpy as np
   import json
   import sys
   import os
   import keras_cv

   directory = "/kaggle/input/practical-innovative-data-science-2024/"
   user_data = directory + "training_data"
   valid_data = directory + "training_data"
   test_data = directory + "label_book/" # this can be the label book, or any other test set you create
   ```

   Notice that all the intermediate output will be stored in the default location "/kaggle/working". You can directly download the data in this directory by the GUI.

2. **How to use GPU?**

   Turn on the accelerator as follows:

   ```
   %pip install -q --upgrade keras-cv

   Note: you may need to restart the kernel to use updated packages.

   import tensorflow as tf
   from tensorflow import keras
   import numpy as np
   import json
   import sys
   import os
   import keras_cv

   directory = "/kaggle/input/practical-innovative-data-science-2024/"
   user_data = directory + "training_data"
   valid_data = directory + "training_data"
   test_data = directory + "label_book/" # this can be the label book, or any other test set you create
   ```

   Session options

   ACCELERATOR
   GPU T4 x2
   Quota: 00:02 / 30 hrs

   LANGUAGE
   Python

   PERSISTENCE
   No persistence

   ENVIRONMENT
   Pin to original environment (2024-10-04)

   The Resources are listed below

- Kaggle GPU: 16G NVIDIA TESLA P100
  - https://www.kaggle.com/docs/efficient-gpu-usage
- Limited to 30+ hrs/week depending on usage.
- Limited to 12hrs/run

3. Run the code in the background

You can also run the code in the background with Kaggle. Firstly, make sure your code is bug-free, as any error in any code block would result in early stopping. Click the "Save Version" button as follows:
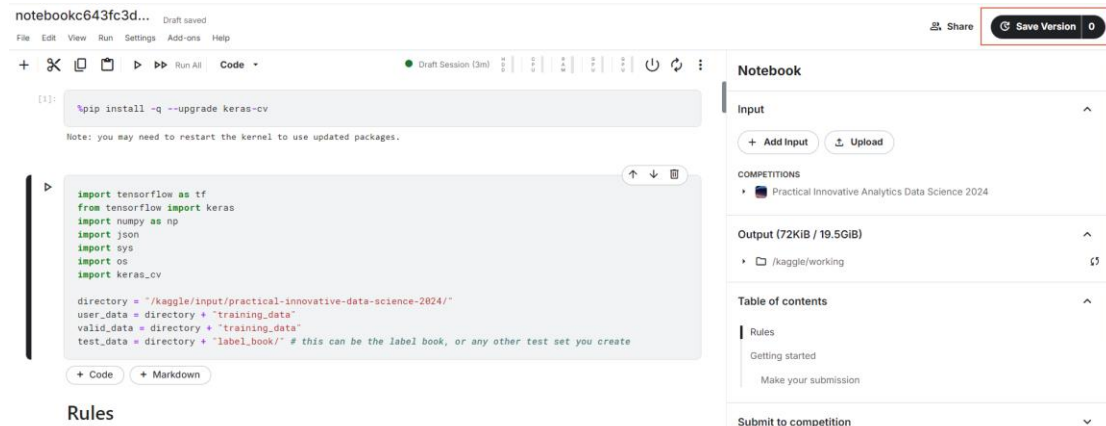
Notice that the output is limited to 20G and the max run time is limited to 12hrs. You can then see the results by clicking the version number:

The log will be shown in the Notebook panel
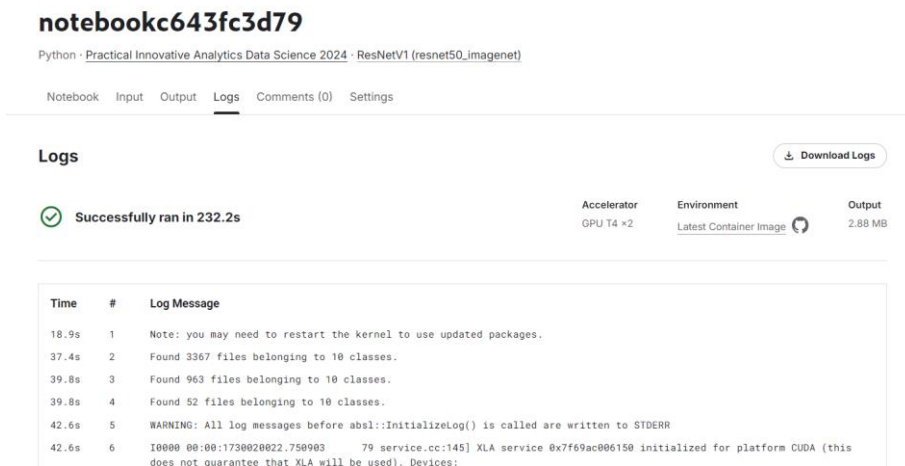
And your output can be accessed through the "Output" panel, where you can

download your data:

Colab tutorial

1. How to get the data?
Register [Kaggle](#) first. In the "Data" panel of the [competition website](#), click Download all:

You can follow the sample notebook first. After exploring, store your data in your google drive and get the shareable link. Execute the following commands:

```
!gdown --fuzzy
https://drive.google.com/file/d/10LEPW3Di4dhqy_NASZ33G0eagi4w3CS4/view?us
p=sharing
```

Remember, DO NOT store input data in your drive and load from there. The input/output is very slow (store at ./ instead). Your output data should be stored in your google drive so that it can be accessed next time.

2. For other tips on using Colab refer to the Colab tutorial at our course website https://phonchi.github.io/nsysu-math608/schedule/.

Tips

1. You can use the sample notebook to prepare your *submission.csv* and Upload your submission to get an estimate of your score
2. Some tips that may help you to improve your accuracy
    - Data cleaning
    - Data preprocessing
    - Fix the label issue
    - Data augmentation
    - Collect or generate more data