# Báo cáo kỹ thuật

## Môi trường thử nghiệm

Chương trình được thử nghiệm trên môi trường Google Colab. Đây là một môi trường notebook Jupyter được host, sử dụng miễn phí và không cần thiết lập trên Google. Với cấu hình như sau:

CPU: Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz

RAM: 8 GB

Chương trình sử dụng các thư viện sau:

Matplotlib

Numpy

Pandas

Sklearn

Transformers

Tensorflow

Và một số thư viện: re, sys, os, …

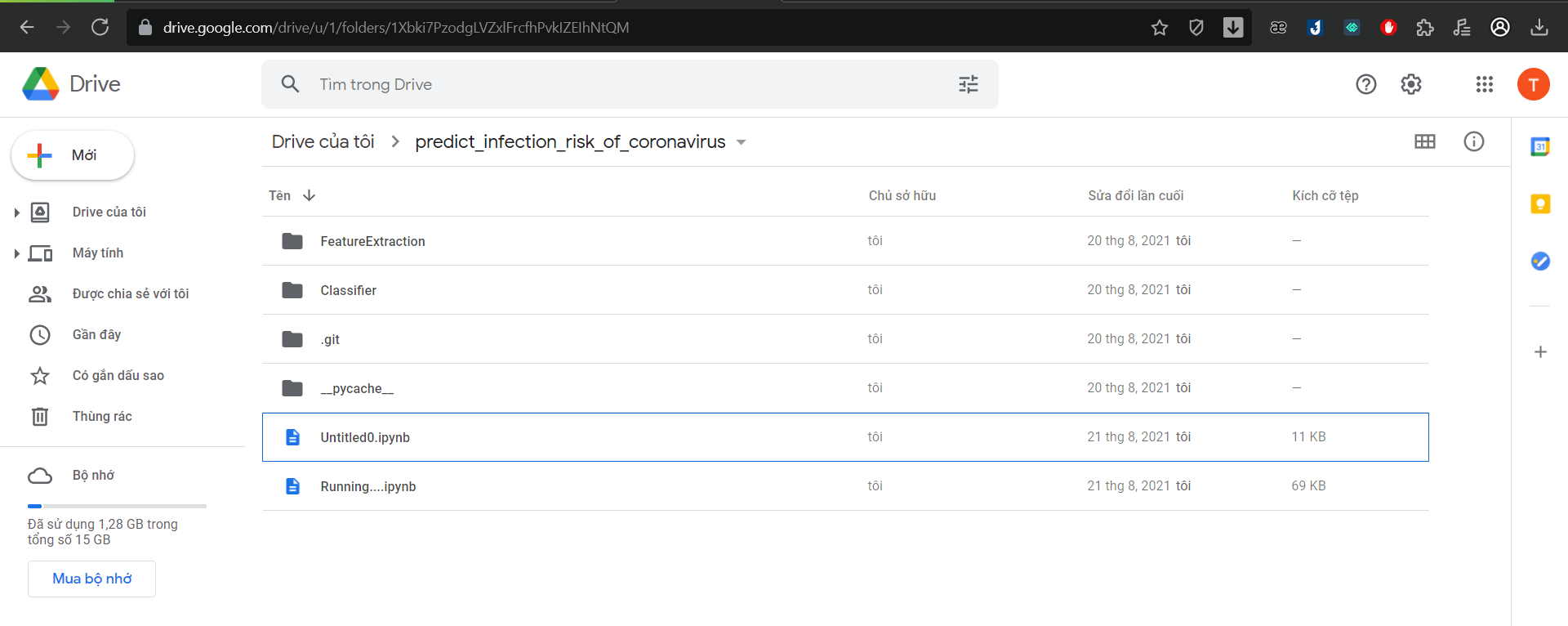
## Cấu trúc thư mục và tên file

## Thư mục “predict\_infection\_risk\_of\_coronavirus”: sử dụng trích xuất đặc trưng AAC, PseAAC và GGAP

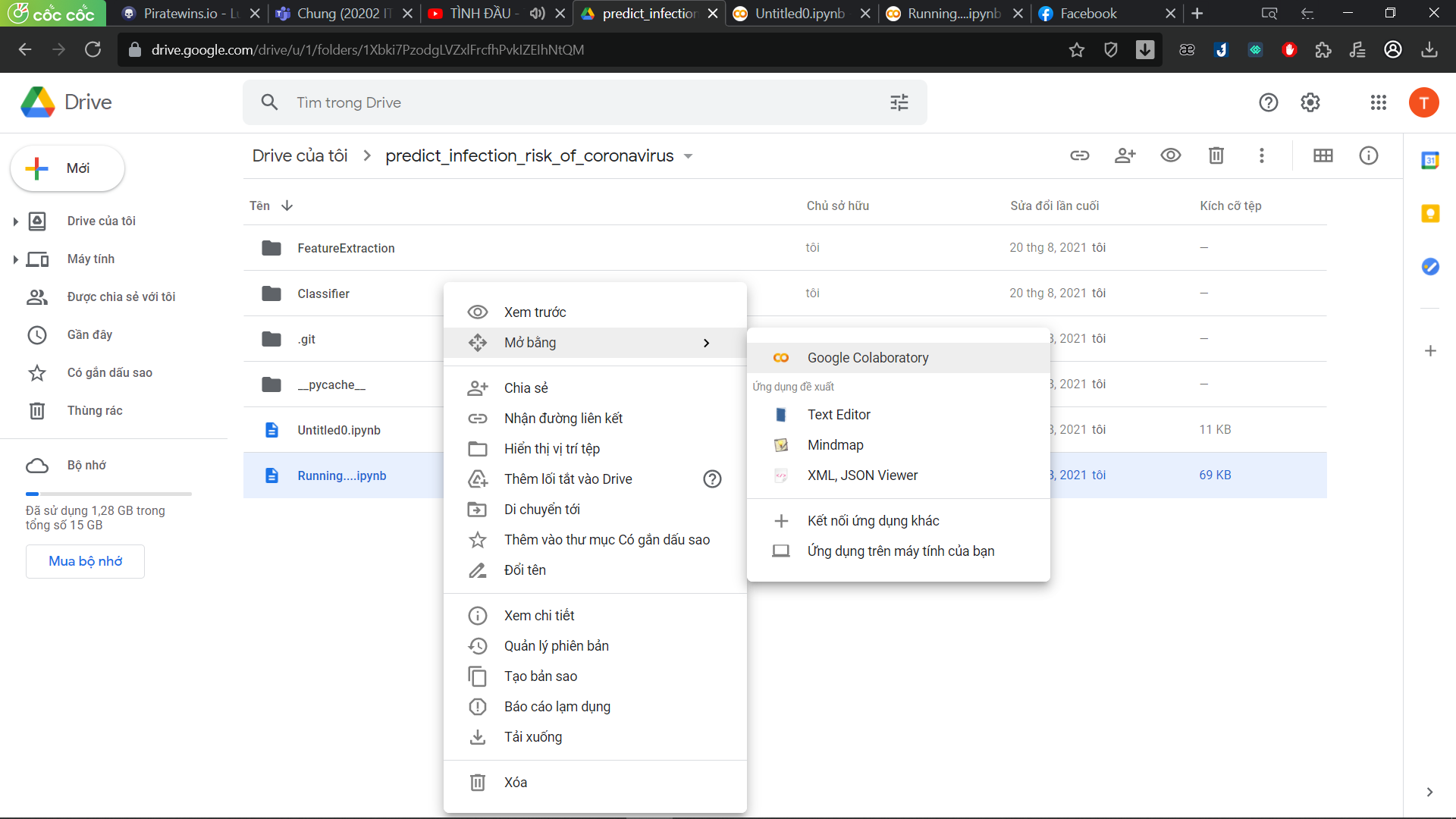
* Thư mục “FeatureExtraction” chứa chương trình để biểu diễn đặc trưng tương ứng với tên thư mục AAC, PseAAC và GGAP. Mỗi thư mục đều có:
* “tên file trùng tên thư mục. py”: chương trình python để trích xuất đặc trưng
* “tên file trùng tên thư mục\_data.csv”: chứa dữ liệu về đặc trưng sau khi trích xuất.
* “tên file trùng tên thư mục\_data\_name.csv”: chứa dữ liệu về đặc trưng sau khi trích xuất kèm thêm cột chứa tên protein.
* “check\_fasta.py”: chương trình kiểm tra file đúng định dạng fasta hay không.
* “readFasta.py” chương trình python sử dụng để đọc chuỗi file định dạng fasta trả về mảng chứa list có 2 thông tin là tên protein và dữ liệu chuỗi protein.
* “data.fasta” chứa dữ liệu về chuỗi protein
* “dev.fasta” chứa dữ liệu về chuỗi protein nhưng với số lượng ít hơn để thuận tiện cho quá trình xây dựng chương trình (giảm thời gian chạy chương trình).
* Thư mục “Classifier” chứa một số file chính:
* “random\_forest\_0.9873817034700315.joblib”: chứa dữ liệu về mô hình rừng ngẫu nhiên.
* “readFasta.py” chương trình python sử dụng để đọc chuỗi file định dạng fasta trả về mảng chứa list có 2 thông tin là tên protein và dữ liệu chuỗi protein.
* “trainning.py”: chương trình python dùng để huấn luyện và kiểm thử.
* “use\_model.py”: Chương trình python sử dụng file dữ liệu mô hình, đọc file fasta và đưa ra dự đoán.
* “summary\_2.txt”: file chứa log kết quả chương trình huấn luyện
* “Running....ipynb” chương trình gọi các chương trình python trong các thư mục con để huấn luyện trên colab

## Cách cài đặt, các log chạy các câu lệnh để cài đặt và thử nghiệm (step-by-step).

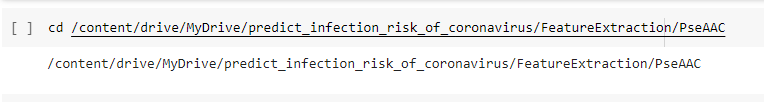
* Bước 1: Tải thư mục lên google drive



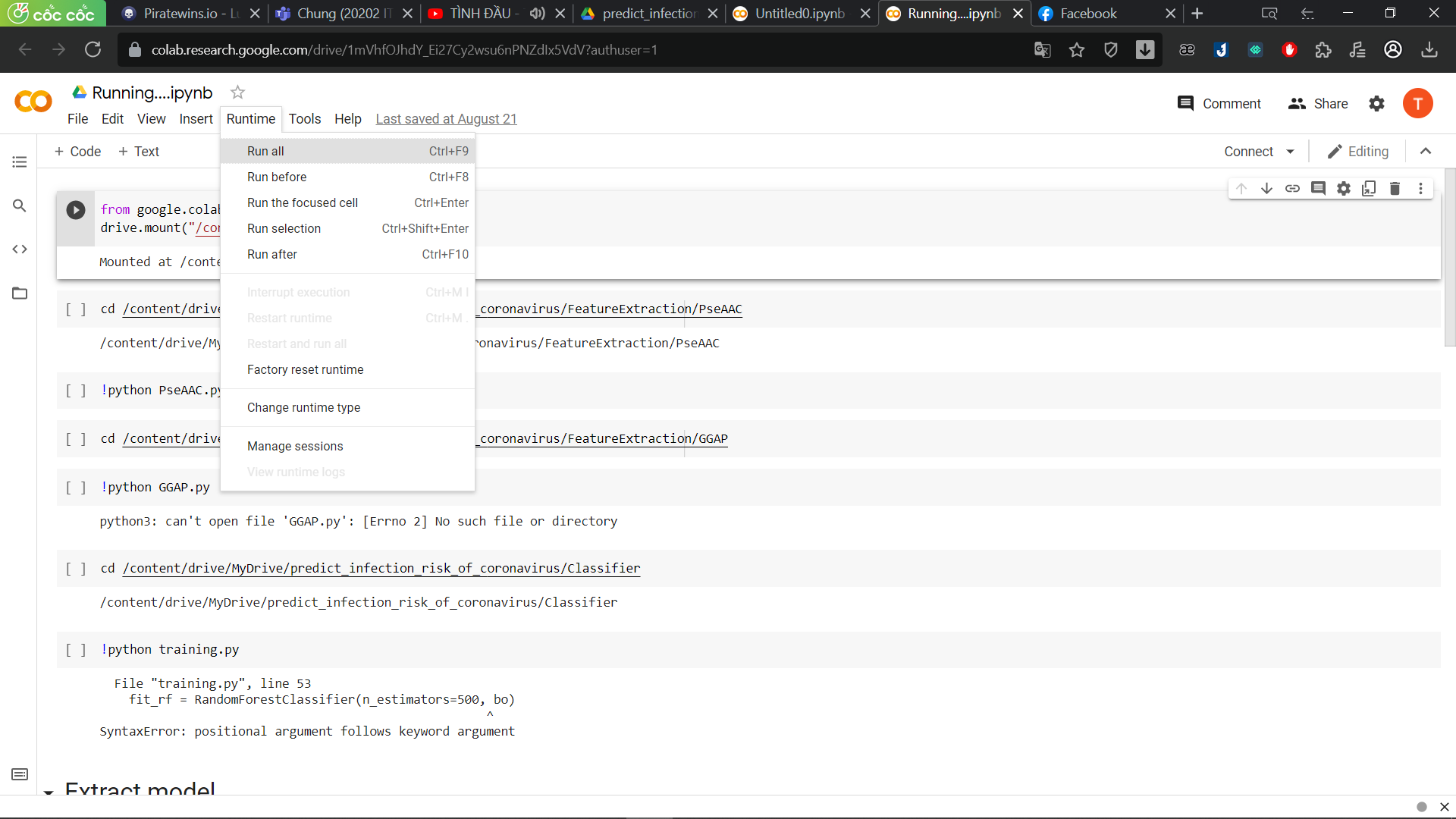
* Bước 2: Sử dụng các file có đuôi “.ipynb” và mở bằng ứng dụng google colab.



* Bước 3: Thay đổi đường dẫn trong các các lệnh liên quan đến di chuyển giữa các folder ứng với vị trí thư mục.



* Bước 4: Thực hiện chạy tất cả chương trình



* Bước 5: Nhận được các log chương trình và kết quả tương ứng.