

Hướng dẫn cách sử dụng hệ thống dự đoán giá xe máy của nhóm 4:

Các thư viện, framework cần cài đặt:

streamlit,
pickle,
pandas,
numpy,
matplotlib,
seaborn,
scikit-learn,
tensorflow,
requests,
beautifulsoup4,
lxml,
xgboost

+ Cấu trúc thư mục của dự án:

```
├── data
│   ├── moto_cleaned.csv
│   ├── moto_cleaned_EDA.csv
│   └── moto_raw.csv
├── model
│   └── model.pkl
├── report
├── slide
└── src
    ├── _1_data_prep_by_
    │   └── web_scrape.py
    ├── _2_data_cleaning
    │   └── data_cleaning.ipynb
    ├── _3_exploratory_data_analysis
    │   └── EDA_moto.ipynb
    ├── _4_build_model
    │   └── model.ipynb
    ├── _5_app
    │   ├── app.py
    │   ├── explore_app.py
    │   ├── predict_app.py
    │   └── __pycache__
    │       ├── explore_app.cpython-310.pyc
    │       ├── predict_app.cpython-310.pyc
    │       └── predict_page.cpython-310.pyc
```

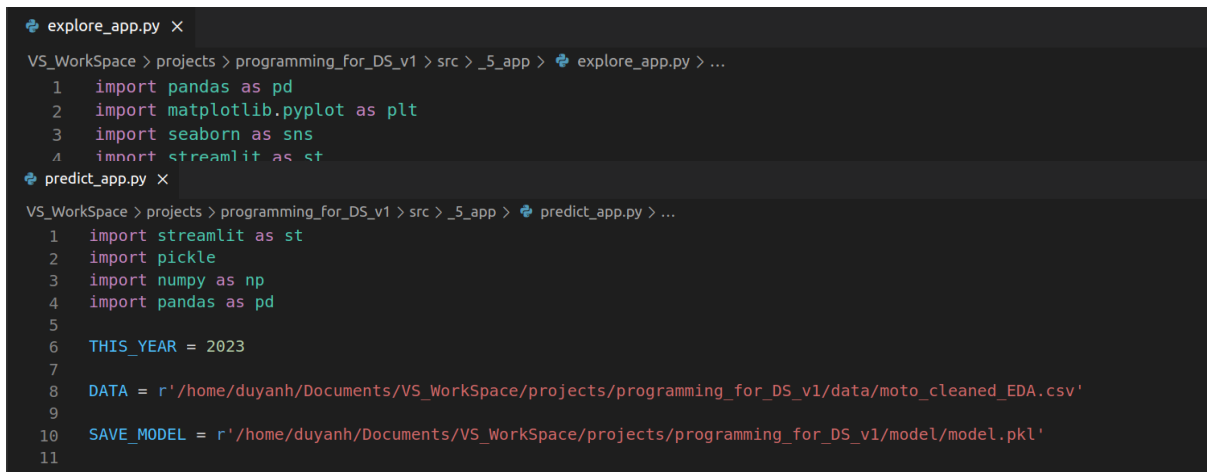
Nếu bạn:

1. Chỉ muốn xem ứng dụng giao diện website của hệ thống:

Bước 1:

Bạn truy cập vào file **explore_app.py** và file **predict_app.py**, sau đó đổi đường dẫn sau tương ứng với thư mục chứa trong máy của bạn

Cụ thể:



```
VS_WorkSpace > projects > programming_for_DS_v1 > src > _5_app > explore_app.py > ...
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import seaborn as sns
4 import streamlit as st

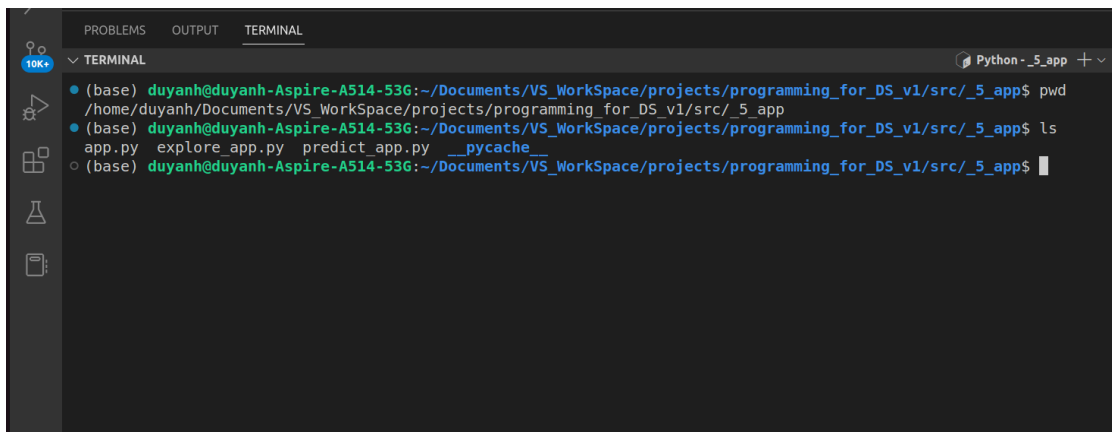
VS_WorkSpace > projects > programming_for_DS_v1 > src > _5_app > predict_app.py > ...
1 import streamlit as st
2 import pickle
3 import numpy as np
4 import pandas as pd
5
6 THIS_YEAR = 2023
7
8 DATA = r'/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/data/moto_cleaned_EDA.csv'
9
10 SAVE_MODEL = r'/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/model/model.pkl'
11
```

Với file **explore_app.py** bạn đổi dòng 6, với file **predict_app.py** bạn đổi hai dòng 8 và 10. Với ba dòng này, bạn chỉ cần đổi đầu

`/home/duyanh/Documents/VS_WorkSpace/projects` về đường dẫn phù hợp với máy của bạn chứa folder `/programming_for_DS_v1`

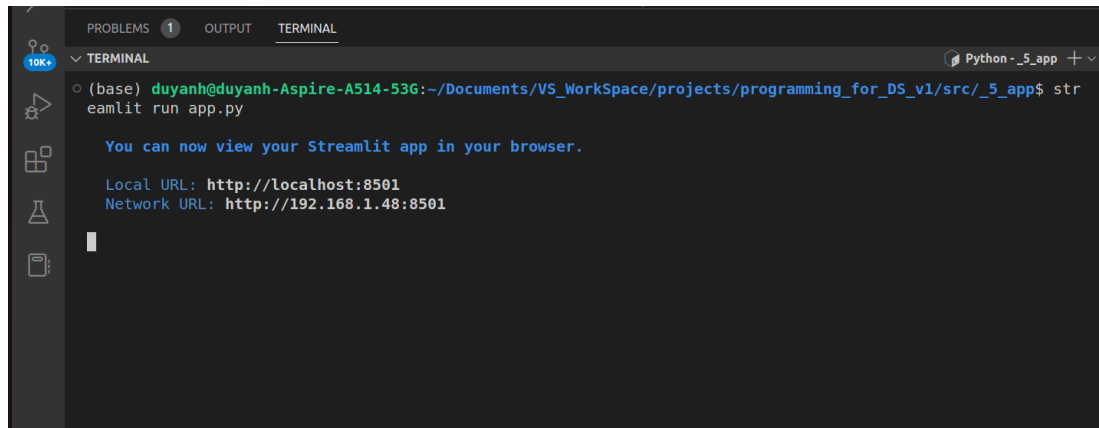
Bước 2:

Trên terminal (bạn có thể dùng VSCode), bạn truy cập đến folder `_5_app`



```
Python - _5_app
TERMINAL
(base) duyanh@duyanh-Aspire-A514-53G:~/Documents/VS_WorkSpace/projects/programming_for_DS_v1/src/_5_app$ pwd
/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/src/_5_app
(base) duyanh@duyanh-Aspire-A514-53G:~/Documents/VS_WorkSpace/projects/programming_for_DS_v1/src/_5_app$ ls
app.py  explore_app.py  predict_app.py  __pycache__
(base) duyanh@duyanh-Aspire-A514-53G:~/Documents/VS_WorkSpace/projects/programming_for_DS_v1/src/_5_app$
```

và chạy lệnh sau: **streamlit run app.py**



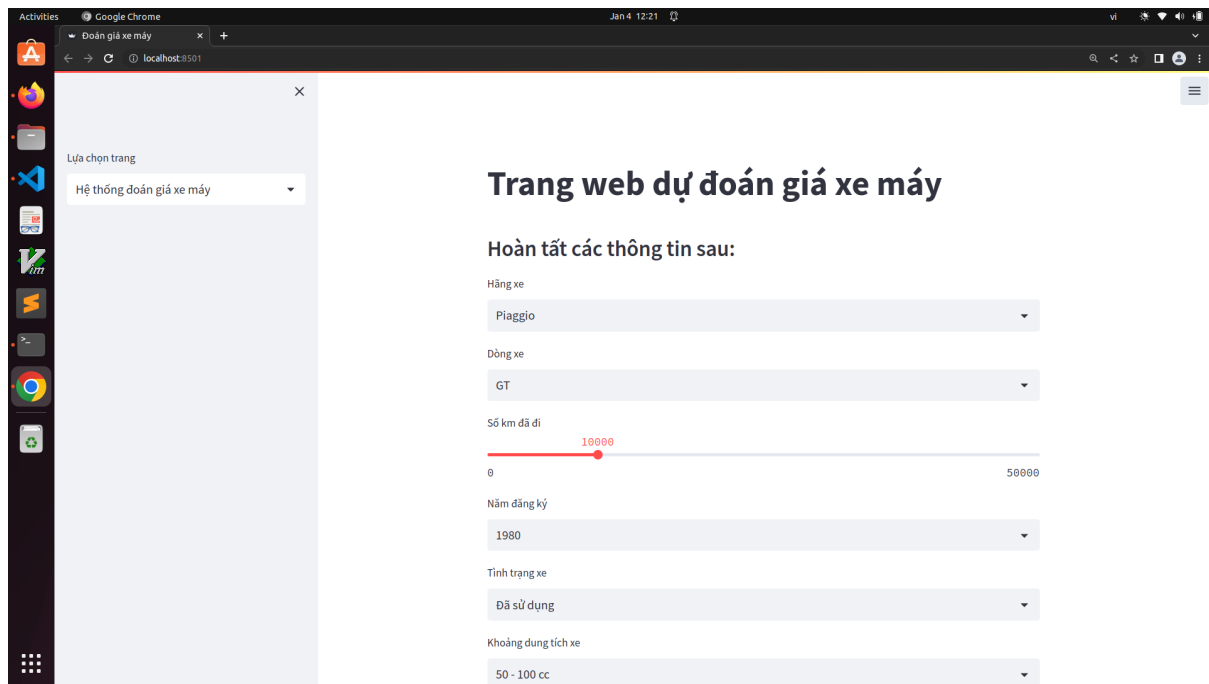
```
(base) duyanh@duyanh-Aspire-A514-536:~/Documents/VS_WorkSpace/projects/programming_for_DS_v1/src/_5_app$ streamlit run app.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://192.168.1.48:8501
```

(Nếu lệnh này không chạy được, bạn hãy thử chạy lệnh sau ***pip install --upgrade protobuf***)

Bạn truy cập vào đường dẫn Local URL trên trình duyệt của bạn và trải nghiệm



2. Muốn chạy thử toàn bộ dự án

Mã nguồn chính của dự án nằm ở thư mục ***/src***

Trong đó có 5 thư mục chính tương ứng với năm quy trình để tạo ra một hệ thống Học máy hoàn chỉnh:

+ Thư mục đầu tiên **/_1_data_prep_by** có chứa file **web_scrape.py** để cào dữ liệu, chuẩn bị dữ liệu cho dự án, bạn có thể chạy thử file này, dữ liệu vào được lưu vào **/data/moto_raw.csv**

```
web_scrape.py X
VS_WorkSpace > projects > programming_for_DS_v1 > src > _1_data_prep_by_ > web_scrape.py > ...
1 from bs4 import BeautifulSoup
2 import requests
3
4
5 MIN_PAGES = 7
6 MAX_PAGES = 250
7 MAIN_DOMAIN = f"https://xe.chotot.com/"
8
9
10 FILE_PATH = r"/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/data/moto_raw.csv"
11
12
13 def check_none(info, main_info_text):
14     if info:
15         return info.text.replace(" ", "")
16     return "Null"
17
```

nhưng bạn cần nhớ đổi đường dẫn như ở mục 1 đối với dòng 10

+ Thư mục thứ 2 **/_2_data_cleaning** có chứa một script notebook (**data_cleaning.ipynb**) để thực hiện quy trình làm sạch dữ liệu, nhưng trước khi bạn chạy và xem kết quả notebook, bạn cũng cần đổi lại đường dẫn, tương tự như những gì đã làm ở mục 1.

```
data_cleaning.ipynb X
1 > src > _2_data_cleaning > data_cleaning.ipynb > M+LÀM SẠCH DỮ LIỆU THÔ (DATA CLEANING) > M+ "Làm sạch" dữ liệu với cột tình trạng xe > M+Ta sẽ thay thế số km đã đi được của các xe này về t
+ Code + Markdown | Run All | Clear Outputs of All Cells | Restart | Variables | Outline ... Python 3
0. Khai báo các thư viện cần thiết

Ta khai báo các thư viện cần thiết sau để phục vụ cho quá trình làm sạch dữ liệu

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Các biến đường dẫn được thay đổi tùy thuộc vào local
PATH = r"/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/data/moto_raw.csv"

SAVE_PATH = r"/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/data/moto_cleaned.csv"

%matplotlib inline

[106] Pyt

1. Tạo dataframe từ dữ liệu thô với các header tương ứng
```

+ Thư mục thứ ba là thư mục **_3_exploratory_data_analysis**, cũng có chứa notebook để phân tích, khám phá dữ liệu, một bước quan trọng để hiểu những đặc trưng của bộ dữ liệu trước khi đưa vào mô hình Học máy, bạn cũng cần chú ý đổi đường dẫn như thư mục 2 nếu muốn chạy notebook.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Các biến đường dẫn được thay đổi tùy thuộc vào local
PATH = r"/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/data/moto_cleaned.csv"

SAVE_PATH = r"/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/data/moto_cleaned_EDA.csv"

df = pd.read_csv(PATH)
df.head(10)
```

+ Thư mục thứ tư là **_4_build_model**, chứa notebook về quá trình xây dựng, tìm ra mô hình phù hợp, những thông tin cần thiết về mô hình được lưu vào file **/model/model.pkl**. Đừng quên đổi tên đường dẫn trước khi chạy nhé

```
from copy import deepcopy

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import SGDRegressor
from sklearn.model_selection import cross_val_score
from sklearn.preprocessing import PolynomialFeatures

from xgboost import XGBRegressor
from sklearn.ensemble import RandomForestRegressor

import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers

import pickle

# Các biến đường dẫn được thay đổi tùy thuộc vào local
PATH = r"/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/data/moto_cleaned_EDA.csv"

SAVE_PATH = r"/home/duyanh/Documents/VS_WorkSpace/projects/programming_for_DS_v1/model/model.pkl"
```