ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO LAB 02:

LÀM VIỆC VÀ TRỰC QUAN HÓA DỮ LIỆU CHUỖI THỜI GIAN

Môn Học: Trực Quan Hóa Dữ Liệu

GVHD: Thầy Bùi Tiến Lên, Cô Nguyễn Thị Thu Hằng, Thầy Nguyễn Bảo Long, Thầy Lê Nhựt Nam

Nhóm: 5

Danh sách sinh viên thực hiện:

1. Đinh Thị Thùy Linh 20120130
2. Lương Vĩnh Phú 20120347
3. Bùi Thanh Tùng 20120398
4. Đỗ Tấn Tài 20120408
5. Trần Khắc Bình 20120437

Thành phố Hồ Chí Minh - 2023

TÓM TẮT

Ô nhiễm môi trường, đặc biệt là ô nhiễm không khí luôn là vấn đế mà cả thế giới quan tâm, nó ảnh hưởng trực tiếp đến sức khỏe con người. Thấy được tầm nghiêm trọng của vấn đề nhóm 5 lựa chọn bộ dữ liệu *U.S. Air Pollution* cho lab 02. Thông qua những phân tích về ô nhiễm không khí mà nhóm đã tìm ra từ bộ dữ liệu, mọi người (những ai đang sống ở Mỹ) có thể có thêm thông tin về lựa chọn các tiểu bang, các khoảng thời gian có thể ra ngoài để ít bị ảnh hưởng đến sức khỏe.

Bộ dữ liệu được chia sẻ miễn phí trên nền tảng data.world (Link dataset: https://data.world/data-society/us-air-pollution-data)

Các bước thực hiện bài lab02:

- Tìm dữ liệu.
- Tiền xử lý dữ liệu: xử lý, làm sạch dữ liệu (các bước trình bày chi tiết được ghi trong file **lab02.ipynb**)
- Trực quan hóa mối quan hệ giữa các trường dữ liệu và rút ra các nhận xét (các bước trình bày chi tiết được ghi trong file **lab02.ipynb**)
- Xây dựng mô hình ARIMA và SARIMA dự đoán chỉ số AQI ở New York. (các bước trình bày chi tiết được ghi trong file **Models.ipynb**)

Cấu trúc file nộp:

- **Thư mục docs:**

  Report.pdf : file báo cáo đồ án
- **Thư mục source_codes:**

  lab02.ipynb : file notebook thực hiện các bước tiền xử lý và phân tích dữ liệu.

  Models.ipynb : file notebook thực hiện xây dựng model.

  requirements.txt: file chứa các thư viện và phiên bản được sử dụng.
- Thư mục **datasets:**

  link_dataset.txt : file chứa link dẫn đến thư mục chứa dataset trên cloud

  pollution_us_2000_2016.csv - tập dữ liệu thô ban đầu, cần tải file này về và lưu vào thư mục **source_codes** để có thể chạy file lab02.ipynb.

  df_cleaned.csv - file dữ liệu sau khi tiền xử lý, KHÔNG CẦN TẢI file này, file sẽ được tạo khi run file lab02.ipynb

Mức độ hoàn thành lab02

STT	Yêu cầu	Công việc	Mức độ hoàn thành
1	Thu thập dữ liệu	Thực hiện theo các yêu cầu phần 2A trong file CSC10108-Lab02.pdf	100%
2	Khám phá dữ liệu	Thực hiện theo các yêu cầu phần 2B trong file CSC10108-Lab02.pdf	100%
3	Khám phá mối quan hệ trong dữ liệu	Trực quan hóa mối quan hệ giữa các trường dữ liệu và rút ra các nhận xét	100%
4	Sử dụng mô hình học máy cơ bản.	Dùng mô hình ARIMA,SARIMA để dự đóa AQI	100%

Bảng phân công công việc

STT	Công việc	Người thực hiện	Nội dung
1	Tìm nguồn dữ liệu, đề tài.	Tùng, Tài, Bình, Linh, Phú	Chọn bộ dữ liệu: pollution_us_2000_2016
2	Tiền xử lý dữ liệu.	Tùng, Tài	Thực hiện như các bước theo file hướng dẫn đồ án, trình bày trong file **lab02.ipynb**
3	Khám phá mối quan hệ giữa các biến dữ liệu	Tùng, Tài, Bình, Linh, Phú	Trực quan hóa mối quan hệ giữa các trường dữ liệu và rút ra các nhận xét
4	Xây dựng mô hình học máy	Tùng	Dùng mô hình ARIMA,SARIMA để dự đóa AQI ở New York
5	Tổng hợp, chỉnh sửa	Tùng, Tài, Bình, Linh, Phú	Tổng hợp, chỉnh sửa file **lab02.ipynb**
6	Viết báo cáo	Tùng, Tài, Bình, Linh, Phú	

Link github: https://github.com/tungbtt/Lab02_DV/

TÀI LIỆU THAM KHẢO

[1]. https://www.kaggle.com/code/prashant111/arima-model-for-time-series-forecasting#10.-Find-the-optimal-ARIMA-model-using-Out-of-Time-Cross-validation-

[2].https://colab.research.google.com/drive/1ebLY9ZAZEKTm7GNL7oCA_8_hpzdeYeWv?authuser=1#scrollTo=deHFPzk0jFVE

[3]. https://www.kaggle.com/code/rohanrao/calculating-aqi-air-quality-index-tutorial#O3-(Ozone-or-Trioxygen)

[4]. https://www.kaggle.com/code/jeffysonar/analysis-of-co-levels-with-plotly/notebook

[5].https://www.kaggle.com/code/fredzanella/sql-tableau-and-forecasting-on-us-pollution-data?fbclid=IwAR18v-aJ73GVORbNbLYQzuK1bJ3Qd48qWW8qZSLOIuMKfAzcFnGs1Fxla2Y#3---Naive-Models

[6].https://colab.research.google.com/drive/1qwsunBsinQRVUX5VTKsrZ-jSh1FAVdsM?authuser=1&fbclid=IwAR2XNGiixoI6BC_1bA-0TMCRfZQe0c1x5DWUgnofk5IXb3GwKpxrujmROfM

[7].https://www.kaggle.com/code/pmw9440/trend-analysis-of-air-pollution-in-u-s-a/?fbclid=IwAR0xHa2VinHpl7cUgD6jIEcbUYXiSlM1A0GzcCKwz-A68kKOvsK3Ofh08Rc