### BỘ GIÁO DỤC VÀ ĐÀO TẠO TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN



KHOA CÔNG NGHỆ THÔNG TIN

**TRỰC QUAN HÓA DỮ LIỆU**

**BÁO CÁO BÀI TẬP MÔN HỌC BÀI TẬP: DV-TW-Lab02**

**Giảng viên hướng dẫn** :Lê Nhựt Nam

**Lớp** :CQ2022/4

### **Nhóm sinh viên thực hiện** : Đinh Viết Lợi- 22120188 Nguyễn Trần Lợi- 22120190 Nguyễn Nhật Long-22120194 Trần Minh Tâm- 22120323

*Hồ Chí Minh, ngày 20 tháng 11 năm 2024*

# MỤC LỤC

Mục lục

[PHẦN 1: BÁO CÁO NHÓM 2](#_bookmark0)

1. [Thông tin thành viên 2](#_bookmark1)
2. [Phân chia công việc 2](#_bookmark2)

[PHẦN 2: TỔNG QUAN ĐỒ ÁN 4](#_bookmark3)

1. [Yêu cầu đồ án 4](#_bookmark4)
   1. [Giới thiệu đồ án 4](#_bookmark5)
   2. [Yêu cầu đồ án 4](#_bookmark6)
2. [Mức độ hoàn thành 4](#_bookmark7)

[PHẦN 3: BÁO CÁO KẾT QUẢ 6](#_bookmark8)

1. [Thu thập dữ liệu 6](#_bookmark9)
2. [Khám phá dữ liệu 6](#_bookmark10)
3. [Trực quan và ý nghĩa 8](#_bookmark11)
   1. [Trực quan dữ liệu 8](#_bookmark12)
   2. [Mô hình dự đoán 9](#_bookmark13)

[PHẦN 4: TÀI LIỆU THAM KHẢO 13](#_bookmark14)

1. [Tài liệu nhóm 13](#_bookmark15)
2. [Tài liệu tham khảo 13](#_bookmark16)

# PHẦN 1: BÁO CÁO NHÓM

## Thông tin thành viên

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Email | Vai trò chính |
| 22120188 | Đinh Viết Lợi | [22120188@student.hcmus.edu.vn](mailto:22120188@student.hcmus.edu.vn) | Trưởng nhóm |
| 22120190 | Nguyễn Trần Lợi | [22120190@student.hcmus.edu.vn](mailto:22120190@student.hcmus.edu.vn) |  |
| 22120194 | Nguyễn Nhật Long | [22120194@student.hcmus.edu.vn](mailto:22120194@student.hcmus.edu.vn) |  |
| 22120323 | Trần Minh Tâm | [22120323@student.hcmus.edu.vn](mailto:22120323@student.hcmus.edu.vn) |  |

## Phân chia công việc

Đồ án được xây dựng và hoàn thành trong khoảng thời gian từ : 20/10/2024 đến 1/11/2024.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nội dung công việc | Người phụ trách | Thời gian bắt đầu | Thời gian kết thúc | Kết quả mong muốn |
| Tìm nguồn dữ liệu |  |  |  |  |
| Quản lý, điều hành nhóm |  |  |  |  |
| Vẽ biểu đồ Radar, Line, Bar Chart |  |  |  |  |
| Vẽ biểu đồ Heat map, Miền, Pie Chart |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Vẽ biểu đồ Box Plot, Histogram, Heatmap |  |  |  |  |
| Vẽ biểu đồ Line, Violin, Venn |  |  |  |  |
| Mô hình học máy |  |  |  |  |
| Làm báo cáo phần 3 |  |  |  |  |
| Làm báo cáo phần 1,2,4 |  |  |  |  |

# PHẦN 2: TỔNG QUAN ĐỒ ÁN

## Yêu cầu đồ án

## Giới thiệu đồ án

* + Đồ án được thực hiện dựa trên yêu cầu của đồ án DV–TW– Lab02 thuộc lớp “Trực quan hóa dữ liệu - CQ2022/4” trường đại học Khoa Học Tự Nhiên thuộc Đại học Quốc gia TP.HCM học kỳ I năm học 2024-2025.
  + Tiếp nối nội dung của đồ án Lab01, đồ án Lab02 xoay quanh các nội dung liên quan đến kĩ năng và công cụ giúp trực quan hóa dữ liệu. Bên cạnh đó là làm quen với kiểu tập dữ liệu time series, phân tích các khái niệm, nội dung cần khai thác đối với kiểu dữ liệu này.
  + Kết quả của đồ án là các mã nguồn được chú thích đầy đủ tên biểu đồ, ý nghĩa, kết luận rút ra từ dữ liệu, mô tả tổng quan về các khoảng thời gian được khai thác. Nội dung của các biểu đồ xoay quanh về chất lượng không khí tại quận Sỹ Lâm (Shilin) thuộc thành phố Đài Bắc, Đài Loan.

## Yêu cầu đồ án

* + Bài làm không có có báo cáo sẽ không được chấm điểm.
  + Các thành viên không đóng góp cho dự án sẽ không nhận được điểm.
  + Các nguồn tham khảo (nếu có) cần được ghi đầy đủ trong báo cáo ở phần Tài liệu tham khảo. Lưu ý rằng cần phân biệt rõ giữa việc tham khảo và đạo văn.
  + Cá nhân hoặc nhóm nào vi phạm gian lận và không trung thực sẽ nhận 0 điểm trong khóa học.
  + Đối với buổi thực hành này, có một số hạn chế cụ thể mà chúng tôi yêu cầu bạn tuân thủ nghiêm ngặt:
    - Không sử dụng các giải pháp phần mềm nâng cao như Tableau cho mục đích trực quan hóa dữ liệu.
    - Nếu muốn sử dụng các thư viện bổ sung, vui lòng tham khảo ý kiến giảng viên trước để nhận được sự chấp thuận.
    - Trong khi việc áp dụng các thuật toán máy học đơn giản có thể cung cấp những hiểu biết sâu sắc hơn về dữ liệu của bạn, đây là yêu cầu tùy chọn và không phải là yêu cầu bắt buộc.

## Mức độ hoàn thành

* **Tổng quan**: nhóm hoàn thành đầy đủ các nội dung cơ bản được yêu cầu và thực hiện tìm hiểu vào các phần nội dung được khuyến khích. Bài làm được đầu tư cẩn thận về chất lượng và số lượng, báo cáo tổng hợp được tất cả quá trình làm việc và kết quả đạt được của nhóm.
* Bộ câu hỏi được chuẩn bị có tính phổ biến và ứng dụng cao, phù hợp với nhu cầu phân tích.
* Nhóm sử dụng đa dạng các loại biểu đồ để trực quan dữ liệu nhằm giải đáp các vấn đề được nêu ra. Tất cả thông tin cần thiết cho việc nắm bắt biểu đồ đều được trình bày rõ ràng, đầy đủ.

 Mức độ hoàn thành: 100%

* **Hạn chế:** nhóm chưa thật sự khai thác được hết tiềm năng của loại biểu đồ time series dẫn tơi việc các biểu đồ không có sự đột phá mạnh mẽ, chỉ xoay quanh mức cơ bản và vận dụng.

# PHẦN 3: BÁO CÁO KẾT QUẢ

## Thu thập dữ liệu

* + Nguồn dữ liệu: Kaggle- [Taiwan Air Quality Index Data 2016~2024](https://www.kaggle.com/datasets/taweilo/taiwan-air-quality-data-20162024/data).
  + Ô nhiễm không khí và bảo vệ khí hậu là một trong những vấn đề được nhiều quốc gia và tổ chức quan tâm, trở thành một trong những chủ đề trọng tâm trong tình hình kinh tế phát triển nhanh chóng tại nhiều quốc gia trong thế kỷ 21. Chất lượng không khí trong những năm gần đây đang có xu hướng trở nên trầm trọng hơn bao giờ hết khi liên tục ghi nhận nhiều khoảng thời gian mức độ chất lượng không khí giảm mạnh và xuất hiện nhiều ca bệnh liên quan tới hô hấp. Nhiều quốc gia và tổ chức đã thực hiện nhiều quan sát, phân tích chất lượng không khí nhằm tìm ra giải pháp hạn chế sự tác động của không khí kém chất lượng.
  + Tập dữ liệu trên được thu thập từ trang web Kaggle, một nền tảng chuyên cung cấp các nguồn tập tài liệu chất lượng với nhiều chủ đề. Tập dữ liệu “ Taiwan Air Quality Index Data 2016-2024” thu thập 24 quan sát một ngày, liên tục từ ngày 25/11/2016 đến 31/8-2024 trên địa bàn của tất cả tỉnh thuộc Đài Loan.
  + Nhằm giảm thiểu sai sót khi phân tích số lượng lớn đồng thời tập trung phân tích chất lượng hơn, nhóm chỉ tập trung phân tích một quận thuộc một tỉnh của Đài Loan. Đối tượng được nhóm quan sát phải đảm bảo mang lại nhiều giá trị từ kết quả phân tích. Từ việc phân tích các thông tin về văn hóa, địa lý, chính trị và kinh tế bên ngoài, nhóm đã quyết định chọn quận Sỹ Lâm (Shilin) thuộc tỉnh Đài Bắc (Taipei). Taipei là một quận hành chính quan trọng, có diện tích lớn và có nhiều địa điểm du lịch thuộc về thành phố Đài Bắc-Thủ đô của Đài Loan.
  + Nguồn dữ liệu trên được sự cho phép tải và nghiên cứu bởi tác giả của nguồn dữ liệu cho mục đích nghiên cứu, tác giả khuyến khích sử dụng nguồn dữ liệu và gợi ý một số cách khai thác tiềm năng của tập dữ liệu.

## Khám phá dữ liệu

* + Bộ dữ liệu được lưu dưới dạng một tập tin csv chứa một bảng thông tin về các quan sát được thực hiện đo đạc về đặc điểm của không khí sau mỗi giờ tại các địa điểm khắp Đài Loan.
  + Mỗi dòng dữ liệu là một quan sát được thực hiện tại một thời gian cụ thể tại một thời điểm cụ thể tại Đài Loan. Mỗi quan sát này sẽ lưu lại các đặc điểm về các chỉ số đặc điểm của một số khí nhất định kèm theo đánh giá tổng quan của người thực hiện về chất lượng không khí tại thời điểm đó. Việc trùng lắp dữ liệu đối với thời gian và địa điểm quan sát sẽ được xử lý trước khi thực hiện phân tích. Các dòng dữ liệu phải cùng thực hiện các quan sát có nghiệp vụ tương tự nhau, các thông số, kiểu dữ liệu phải phù hợp với tất cả các dòng xung quanh.
  + Nếu tồn tại các dòng dữ liệu nhiễu hoặc dữ liệu không phù hợp với cột dữ liệu tương ứng sẽ gây khó khăn cho việc phân tích và loại bỏ các giá trị này để đảm bảo dữ liệu phù hợp trước khi phân tích.
  + Bộ dữ liệu bao gồm một tập thuộc tính đa dạng về cả giá trị lẫn kiểu dữ liệu, cung cấp cho người sử dụng một nguồn thông tin đáng giá và cái nhìn tổng quan về vấn đề khi thực hiện phân tích. Các thuộc tính của dữ liệu và kiểu dữ liệu của chúng:
    - date- Chuỗi (object): thời điểm thực hiện quan sát chi tiết đến năm tháng ngày giờ.
    - sitename- Chuỗi (object): quận của thành phố.
    - county- Chuỗi (object): thành phố thuộc Đài Loan.
    - aqi- Số nguyên (int): chỉ số chất lượng không khí air quality index.
    - pollutant- Chuỗi (object): tình trạng ô nhiễm chính.
    - status- Chuỗi (object): đánh giá chất lượng không khí theo cấp độ.
    - so2- Số nguyên (float): nồng độ khí SO2, đơn vị ppb (parts per billion).
    - co- Số nguyên (float): nồng độ khí CO, đơn vị ppm (parts per million).
    - o3- Số nguyên (float): nồng độ khí O3, đơn vị ppb (parts per billion).
    - O3\_8hr: nồng đồ khí O3 trung bình đo được sau 8 giờ liên tiếp.
    - pm10: nồng độ các hạt bụi mịn có đường kính nhỏ hơn 10 µm, đơn vị μg/m3.
    - pm2.5: nồng độ các hạt bụi mịn có đường kính nhỏ hơn 2.5 µm, đơn vị μg/m3
    - no2- Số nguyên (float): nồng độ khí NO2, đơn vị ppb (parts per billion).
    - nox- Số nguyên (float): nồng độ khí NOX, đơn vị ppb (parts per billion).
    - NO- Số nguyên (float): nồng độ khí NO, đơn vị ppb (parts per billion).
  + Nhìn chung hầu như tất cả các thuộc tính đều có kiểu dữ liệu phù hợp với nhu cầu hiểu về dữ liệu với người sử dụng. Việc quy chuẩn một số thuộc tính như stress, Alcohol Consumption, diet vào các danh sách phân chia theo mức độ giúp dễ dàng hơn trong việc phân nhóm dữ liệu. Tuy nhiên điều này cũng có thể mang lại tác hại nếu cách phân loại của người tạo nên dữ liệu khác với người phân tích, nhiều khi ta không thể biết điều kiện, lí do và các nguyên nhân tạo

nên giá trị dữ liệu này.

* + Một số thuộc tính sẽ không được sử dụng vì không có tính khách quan như chỉ số BMI, Medication Use.
  + Nhóm chỉ thực hiện việc tiền xử lý dữ liệu đối với cột huyết áp của bệnh nhân. Hiện tại do cột dữ liệu này lưu giá trị huyết áp dưới dạng chuỗi “huyết áp tâm thu/huyết áp tâm trương”, việc này tuy đúng với các ghi chép thực tế nhưng không phù hợp với việc trực quan hay phân tích dữ liệu do đó sẽ cần được tách ra làm hai cột thuộc tính riêng là Systolic blood pressure và Diastolic blood pressure. Ta chỉ cần lấy ra các giá trị của cột Blood Pressure, thực hiện phân tích cú pháp chuỗi để lấy ra 2 danh sách giá trị huyết áp tương ứng và kết hợp lại với bảng dữ liệu ban đầu.

## Trực quan và ý nghĩa

## Trực quan dữ liệu

* + Bộ dữ liệu sau khi tiền xử lý tuy có rất nhiều quan sát và thuộc tính giúp mang lại rất nhiều trực quan có thể thực hiện tuy nhiên nhóm chỉ lựa chọn một số biểu đồ hỗ trợ trực quan các vấn đề đáng quan tâm. Các vấn đề nhóm cho rằng quan trọng:
    - **Các thói quen sinh hoạt, bệnh lý nào sẽ gây nên nguy cơ nhồi máu cơ tim ?** : Nhóm phân tích xem trong 5 yếu tố tiểu đường, tiền sử người thân, hút thuốc, béo phì, tiêu thụ rượu bia thì yếu tố nào là tác nhân quan trọng nhất. Biểu đồ Radar được sử dụng trong trường hợp này.
    - **Huyết áp của một cá nhân bị ảnh hưởng thế nào bởi thời gian họ ngủ ?** : Tình trạng thiếu ngủ do làm việc quá sức, thức đêm giải trí của giới trẻ dẫn đến việc thiếu ngủ có phải là nguyên nhân khiến gia tăng nguy cơ nhồi máu cơ tim hay không. Biểu đồ đường sẽ thể hiện mức huyết áp trung bình của các bệnh nhân có chung khoảng thời gian ngủ.
    - **Độ tuổi là nạn nhân phổ biến của nhồi màu cơ tim ?** : Việc biết được câu trả lời không chỉ giúp đánh giá tình hình mắc bệnh ở các độ tuổi mà đồng thời còn giúp cho người xem cân nhắc điều chỉnh thói quen sinh hoạt sao cho phù hợp với độ tuổi của mình dựa trên nguy cơ mắc ở độ tuổi này. Biểu đồ tròn sẽ giúp thể hiện tổng quan mức độ của từng độ tuổi so với tổng thể.
    - **Liệu việc tập thể dục thường xuyên có thực sự giảm nguy cơ nhồi máu cơ tim ?:** Nhóm muốn tìm hiểu xem liệu việc có thói quen luyện tập thể dục

đều đặn và chất lượng liệu có giúp con người tránh nguy cơ nhồi máu cơ tim hay không và liệu có áp dụng cho mọi độ tuổi. Biểu đồ heatmap sẽ cho cái nhìn tổng quan và so sánh cụ thể trong trường hợp này.

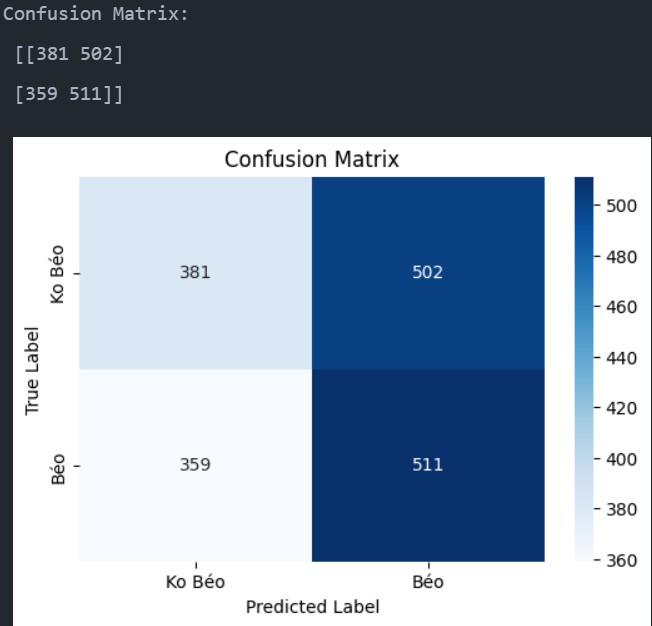
* + - **Liệu việc ngồi một chỗ trong thời gian dài có tăng nguy cơ nhồi máu cơ tim hay không ?:** Bên cạnh mục đích kiểm tra xem thói quen ít vận động của từng nhóm tuổi , ta đồng thời có thể xem liệu việc ngồi nhiều ít ra ngoài của con người có gây nên nguy cơ mắc bệnh không. Biểu đồ miền sẽ phù

hợp trong trường hợp này.

* + - **Mức độ stress của từng nhóm tuổi và ảnh hưởng đến nguy cơ nhồi màu cơ tim:** Stress là một từ khóa xuất hiện rất phổ biến trong nhiều năm trở lại đây, đặc biệt là ở người trẻ do nhiều vấn đề từ cuộc sống. Nhóm muốn kiểm chứng xem độ tuổi nào đang chịu ảnh hưởng nặng nề nhất từ stress và số lượng người mắc stress cao ( từ 6 trở lên). Và đặc biệt hơn là liệu đây có phải là một nhân tố quan trọng thầm lặng gây nên nhồi máu cơ tim. Biểu đồ heatmap sẽ cho cái nhìn toàn thể về hiện trạng này.
    - **Tiểu đường có đang trở thành căn bệnh chung của con người? :**Nhiều năm trở lại đây, do sự dày đặt của công việc và công nghệ, càng nhiều người trở nên mờ nhạt với các hoạt động thể chất từ trẻ nhỏ đến người già. Nhóm muốn kiểm tra xem ở độ tuổi nào thì bắt đầu có sự xuất hiện của bệnh béo phì. Biểu đồ tần suất (histogram) sẽ cho thấy sự phân bố của tập dữ liệu theo từng nhóm tuổi của các bệnh nhân mắc bệnh béo phì.
    - **Nhồi máu cơ tim có bị ảnh hưởng bởi vấn đề địa lý hay không và các bệnh nhân ở châu lục này sinh hoạt thế nào? :** Nhóm bên cạnh việc tham khảo xem môi trường sinh sống của các châu lục này có ảnh hưởng đến nhồi máu cơ tim hay không và thói quen sinh hoạt của những cá nhân ở vùng này như thế nào. **Biểu đồ Boxplot** được sử dụng trong trường hợp này.
    - **Mức độ triglycerides có tăng nguy cơ nhồi máu cơ tim hay không ? :** triglycerides là một chất dinh dưỡng cần thiết và phổ biến trong bữa ăn của con người tuy nhiên gần đây lại bị lạm dụng tiêu thụ quá mức làm tăng mức độ chất này có trong máu. Nhóm muốn kiểm tra xem tác động của chất này đến nguy cơ mắc nhồi màu cơ tim của hai giới tính nam và nữ. Biểu đồ Violin được sử dụng để có cái nhìn tổng quát về hai giới tính và hai nhóm có hay không mắc bệnh.
    - **Tiểu đường phổ biến như thế nào?** tương tự với chất béo, đường là một thực phẩm còn được sử dụng nhiều và rộng rãi hơn với mọi độ tuổi. Trước khi kiểm tra tiểu đường ảnh hưởng thế nào đến nhồi máu cơ tim, ta cần kiểm tra xem độ tuổi nào đang mắc bệnh tiểu đường nhiều nhất. Biểu đồ đường rất phù hợp cho việc quan sát sự thay đổi theo từng nhóm tuổi.
    - **Lối sống nhiều thói quen xấu ảnh hưởng thế nào đến nguy cơ mắc nhồi máu cơ tim? :** không chỉ kiểm thử tác động của từng thói quen đến nguy cơ mắc bệnh, ta cần xem thử việc liệu một người có nhiều hơn 1 thói quen xấu sẽ dẫn tới nguy cơ mắc bệnh như thế nào. Biểu đồ Veen rất phù hợp để xem giao thoa giữa các thói quen này, thể hiện bởi số lượng bệnh nhân mắc bệnh.
  + Chi tiết về các biểu đồ, loại biểu đồ, ý nghĩa, nguyên nhân sử dụng và kết luận của từng biểu đồ được mô tả tại các file ipynb kèm giải thích của mã nguồn.

## Mô hình dự đoán

* + ewr



* + **Kết luận:** Mô hình hồi quy logistic này được thiết kế để ‘**dự đoán khả năng một người có bị béo phì hay không’** dựa trên các đặc trưng đầu vào từ dữ liệu. Cụ thể, mô hình sẽ dự đoán nhãn đầu ra là:
    - "Béo” nếu mô hình cho rằng khả năng người đó bị béo phì là cao.
    - "Không béo" nếu mô hình cho rằng người đó không bị béo phì.
  + **Ý nghĩa kết quả dự đoán:** Kết quả dự đoán từ mô hình giúp xác định khả năng béo phì của một người dựa trên các yếu tố đầu vào như chế độ ăn uống, giới tính, quốc gia, lục địa, và bán cầu sinh sống. Điều này có thể hỗ trợ trong các ứng dụng y tế và sức khỏe cộng đồng, cho phép xác định các nhóm đối tượng có nguy cơ cao và đưa ra các biện pháp phòng ngừa béo phì hiệu quả.

# PHẦN 4: TÀI LIỆU THAM KHẢO

## Tài liệu nhóm

1. Đường dẫn github: https://github.com/Dzivilord/DV-TW-Lab02.git
2. Tài liệu thảo luận: [Question+ Problem - Google Docs](https://docs.google.com/document/d/10Eh42hGbU4VzgAQK05JxX7XmI_JP_mnNCRXoMtbgu_w/edit?tab=t.95r34xnhazwq)

## Tài liệu tham khảo

1. Python Data Visualization Cookbook.
2. Aurélien Géron, ‘Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow’.
3. Kyle Gallatin & Chris Albon, ‘Machine Learning with Python Cookbook’.
4. ProgrammingKnowleage, ‘Matplotlib Tutorial for Begineers (Python)|Learn Data…’ [Trực tuyến]. Địa chỉ: https://[www.youtube.com/playlist?list=PLS1QulWo1RIZ3tcrdZodjuXTDTIIXH8](http://www.youtube.com/playlist?list=PLS1QulWo1RIZ3tcrdZodjuXTDTIIXH8) EW
5. GeeksforGeeks, ‘Time Series Analysis & Visualization in Python’ [Trực tuyến]. Địa chỉ: https://www.geeksforgeeks.org/time-series-data-visualization-in-python/