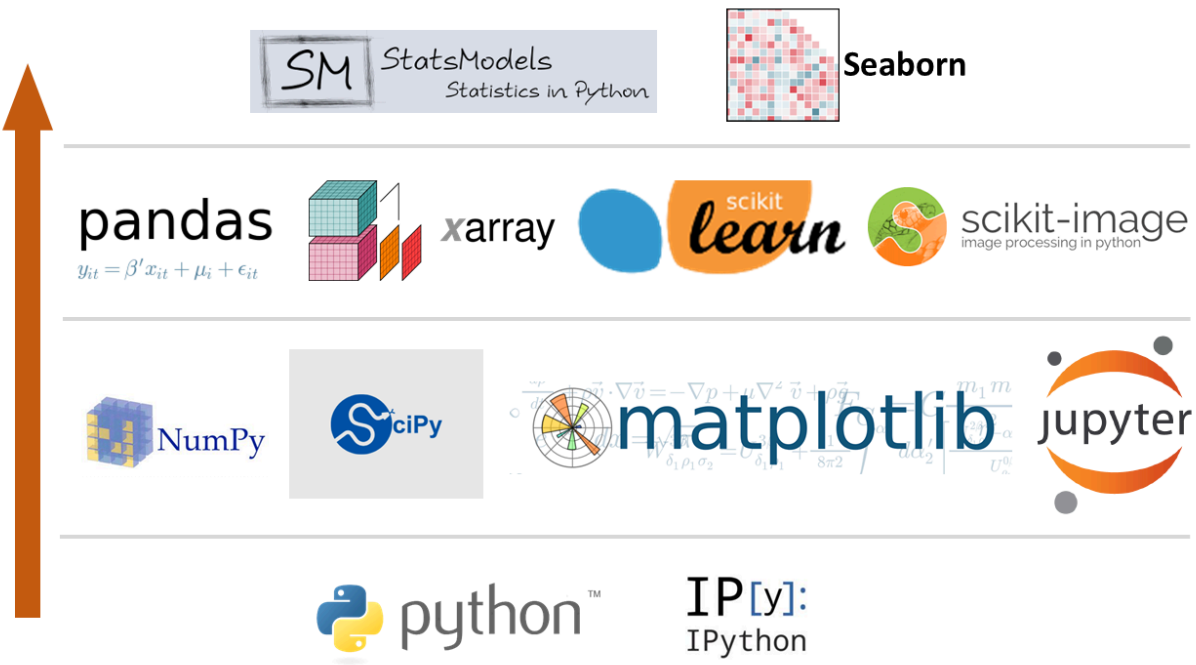
# HOMEWORK 2: NUMPY FOR DATA SCIENCE
# CSC17104 - PROGRAMMING FOR DATA SCIENCE

## 1. TỔNG QUAN VỀ BÀI TẬP

Numpy hay Numerical Python là một thư viện hỗ trợ tính toán số học được sử dụng rộng rãi trong lập trình Python. Nó hỗ trợ thao tác với cấu trúc mảng đa chiều lớn rất tiện lợi và đảm bảo hiệu suất cao. Tìm hiểu về thư viện này là một bước rất quan trọng trong việc học Khoa học Dữ liệu nói riêng và ứng dụng Khoa học Máy tính nói chung bởi vì nó là một "module quan trọng" và là nền tảng của nhiều thư viện mạnh trong bộ công cụ Khoa học Dữ liệu của Python (Pandas, Matplotlib, Scikit-learn).



Trong bài tập này, sinh viên sẽ được tiếp xúc và trải nghiệm thư viện Numpy thông qua việc làm một mini-project. Bên cạnh đó, sinh viên cũng được khuyến khích sử dụng các thư viện hỗ trợ trực quan hoá kết quả để minh hoạ thêm cho project của mình.

## 2. MỤC TIÊU CỦA BÀI TẬP

Sau khi hoàn thành bài tập này, sinh viên sẽ có khả năng:

- Sử dụng tương đối thành thạo Numpy để xử lý và tính toán với dữ liệu dạng bảng.
- Thực hiện đặt một số câu hỏi để hiểu thêm về bộ dữ liệu và trả lời bằng dữ liệu thông qua trực quan hoá đơn giản hoặc sử dụng mô hình học máy đơn giản.
- Sử dụng một số thư viện hỗ trợ trực quan hoá dữ liệu như Matplotlib, Seaborn để minh hoạ kết quả.
- Cài đặt từ đầu một số mô hình học máy đơn giản bằng Numpy như Linear Regression, Logistic Regression, Naive Bayes để huấn luyện và đưa ra kết quả dự đoán cho dữ liệu mới. (phần nâng cao, sinh viên hoàn thành có thể đạt được 9+).

## 3. YÊU CẦU CỦA BÀI TẬP

Sinh viên chọn **một** trong các bài toán sau để thực hiện:

| Bài toán | Đầu vào | Đầu ra | Dữ liệu |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Phát hiện gian lận thẻ tín dụng trong các giao dịch | Thông tin giao dịch thẻ tín dụng | Giao dịch có/không có gian lận | Credit Card Fraud Detection |
| Dự đoán khả năng có thay đổi công việc Khoa học dữ liệu | Các thông tin cơ bản của một ứng viên | Có/Không muốn thay đổi việc công việc | HR Analytics: Job Change of Data Scientists |
| Khảo sát sự đón nhận các loại vacxin | Dữ liệu về Pfizer & BioNTech vaccine. | Sự đón nhận/không đón nhận của công chúng | Pfizer Vaccine Tweets |
| Xác định khả năng rời đi của khách hàng dựa trên chỉ số tín dụng | Thông tin khách hàng tín dụng | Tài khoản của khách hàng đã bị đóng/còn tồn tại | Credit Card customers |
| Phân tích dữ liệu Airbnb | Hoạt động của chủ và khách | Khám phá, trực quan, dự đoán | New York City Airbnb Open |

| | | | |
|---|---|---|---|
| | | về dữ liệu Airbnb | Data |
| Xây dựng hệ thống gợi ý sản phẩm/phim | User ID và lịch sử đánh giá/tương tác của người dùng | Nhận xét dữ liệu, dự đoán danh sách sản phẩm/phim gợi ý cho người dùng | Amazon - Ratings (Beauty Products) |

**Về yêu cầu phân tích dữ liệu**

Sinh viên đặt một số câu hỏi để hiểu về dữ liệu và đưa ra những phân tích, nhận định hoặc khuyến nghị cho người đọc.

Ví dụ khi xem xét dữ liệu Airbnb:

- Chúng ta có thể tìm hiểu gì về các chủ sở hữu và khách hàng khác nhau?
- Chúng ta có thể thấy được gì từ dự đoán? (ví dụ: địa điểm, giá cả, đánh giá, v.v.)
- Những chủ sở nào “bận rộn” nhất và tại sao?
- Có sự khác biệt đáng chú ý nào về lưu lượng truy cập giữa các khu vực khác nhau không và nguyên nhân có thể là gì?

**Về yêu cầu kỹ thuật**

1. Xử lý dữ liệu và tính toán
   - **CHỈ** sử dụng thư viện NumPy cho tất cả các tác vụ:
     - Đọc và load dữ liệu
     - *Tiền xử lý dữ liệu*: kiểm tra tính hợp lệ của giá trị; xác định và loại bỏ (nếu cần thật sự cần thiết) đối với các giá trị ngoại lai; chuẩn hoá (Normalization) cho từng đặc trưng (min-max, log transformation, decimal scaling) trong trường hợp ; điều chuẩn khoảng giá trị phù hợp dữ liệu Unknown hoặc Non-Gaussian Distribution; Điều chuẩn dữ liệu (Standardization) hay z-score để đạt trung bình 0 và phương sai 1 trước khi sử dụng thuật toán dựa trên gradient hoặc feature engineering bằng các kỹ thuật dimensional reduction.
     - *Xử lý missing values* (filling missing values bằng cách sử dụng mean/ median/ specific values; building một model predictor để dự đoán giá trị trị bị thiếu)
     - *Feature engineering* (tạo thêm những đặc trưng mới cho dữ liệu, ví dụ như: mối liên hệ giữa thời gian và rating của người dùng với một sản phẩm bán trên Amazon)
     - Sử dụng một số phép tính số học trong việc tối ưu số học, hạn chế sai số trong quá trình tính toán; tính toán thống kê mô tả, kiểm định giả thiết thống kê (nêu rõ giả thiết H0, giả thiết thay thế H1).

- **Không được phép** sử dụng Pandas, Scikit-learn hay các thư viện xử lý dữ liệu khác.

2. Trực quan hoá dữ liệu đơn giản
   - **CHỈ** sử dụng **Matplotlib** và **Seaborn** để:
     - Vẽ biểu đồ phân tích dữ liệu (histogram, scatter plot, heatmap, v.v.) để minh hoạ kết quả cho các phần trả lời câu hỏi
     - Giải thích xu hướng bằng các biểu đồ đường (line chart)
     - Giải thích tỷ lệ phần trăm bằng các biểu đồ tròn
     - Giải thích bằng cách kết hợp nhiều biểu đồ khác nhau.
   - **Không được phép** sử dụng Plotly, Bokeh hay các thư viện visualization khác
3. Xây dựng Mô hình (nếu có)
   - Được phép sử dụng Scikit-learn.
   - **NẾU** sử dụng **NumPy** để implement lại từ đầu **ĐỂ**:
     - Thuật toán học máy (Linear Regression, Logistic Regression, KNN, v.v.)
     - Hàm mất mát và các thuật toán tối ưu
     - Các độ đo đánh giá (accuracy, precision, recall, F1-score, v.v.)
     - Cross-validation

THÌ sẽ có điểm cao hơn.

- **Không được phép** sử dụng các framework ML/DL như TensorFlow, PyTorch

**Yêu cầu về Repository GitHub**

Cấu trúc thư mục:

```
project-name/
├── README.md
├── requirements.txt
├── data/                  # Nếu dữ liệu quá lớn, thì chỉ cần đính kèm link trong file .txt
│   ├── raw/               # Dữ liệu gốc
│   └── processed/         # Dữ liệu đã xử lý
├── notebooks/
│   ├── 01_data_exploration.ipynb
│   ├── 02_preprocessing.ipynb
│   └── 03_modeling.ipynb
├── src/
│   ├── __init__.py
│   ├── data_processing.py
│   ├── visualization.py
│   └── models.py
```

**README.md phải bao gồm:**

1. **Tiêu đề và Mô tả ngắn gọn** về project
2. **Mục lục**
3. **Giới thiệu:** Mô tả bài toán; Động lực và ứng dụng thực tế; Mục tiêu cụ thể
4. **Dataset:** Nguồn dữ liệu; Mô tả các features; Kích thước và đặc điểm dữ liệu
5. **Method:** Quy trình xử lý dữ liệu; Thuật toán sử dụng (với công thức toán học); Giải thích cách implement bằng NumPy (nếu có)
6. **Installation & Setup**
7. **Usage:** Hướng dẫn cách chạy từng phần
8. **Results:** Kết quả đạt được (metrics); hình ảnh trực quan hoá kết quả thông qua biểu đồ; So sánh và phân tích
9. **Project Structure:** Giải thích chức năng từng file/folder
10. **Challenges & Solutions:** Khó khăn gặp phải khi dùng NumPy; Cách giải quyết
11. **Future Improvements:** Hướng phát triển tiếp theo
12. **Contributors**
    - Thông tin tác giả
    - Contact
13. **License**

## 4. TIÊU CHÍ ĐÁNH GIÁ

| Tiêu chí | Nội dung đánh giá | Tỷ lệ điểm |
|---|---|---|
| 1. Trình bày notebook & Tổ chức mã nguồn | **1.1 Notebook Presentation (10%):** Cấu trúc rõ ràng, Giải thích chi tiết (5%); Visualizations (5%)<br>**1.2 GitHub Repository (10%):** README.md chất lượng (5%); Cấu trúc repository hợp lý, Code organization (5%) | 20% |
| 2. Kỹ thuật NumPy & Optimization | **2.1 Vectorization (10%):** KHÔNG dùng for loops cho operations trên arrays; Sử dụng broadcasting hiệu quả; Áp dụng universal functions (ufuncs)<br>**2.2 Numpy Techniques (15%)**: Fancy indexing & masking; Array manipulation (reshape, transpose, stack, etc.); Memory-efficient operations; Sử dụng np.einsum cho phép | 50% |

| | | |
|---|---|---|
| | tính phức tạp<br>**2.3 Mathematical operations (10%):** Tính toán số học, ổn định số học, hạn chế độ lỗi.<br>**2.4 Code Efficiency (15%):** Clean, hiệu quả, hạn chế tài nguyên, có khả năng tái thực nghiệm lại. | |
| 3. Kết quả phân tích | **3.1 Model Performance (15%)**<br>**3.2 Analysis & Insights (15%)** | 30% |
| 4. Bonus | **4.1** Bài làm có những phân tích thú vị;<br>**4.2** Cài đặt thuật toán học máy từ đầu mà chỉ dùng Numpy. | Tối đa 10% |

Nộp bài:

- Sinh viên nén thư mục nộp thành một file duy nhất và đặt tên với định dạng: <MSSV>.zip. Ví dụ: 2312001.zip
- Nếu file quá 25MB, sinh viên nộp link bài làm đã được upload qua Onedrive/ Google Drive/ Github và nộp qua Moodle bằng file <MSSV>.txt. Ví dụ: 2312001.txt

## 5. LƯU Ý

- Đây là bài tập cá nhân.
- Thời lượng: 3 tuần.
- Mọi hành vi đạo văn, gian lận, hoặc gian dối sẽ bị 0 điểm toàn môn học.

## 6. THÔNG TIN LIÊN HỆ

Nếu bạn có bất kỳ câu hỏi nào về dự án này, vui lòng liên hệ với giảng viên theo địa chỉ:

Lê Nhựt Nam: lnnam@fit.hcmus.edu.vn (*Giảng viên sẽ trả lời câu hỏi của bạn sớm nhất có thể.*)