GIỚI THIỆU MÔN HỌC

Course: Nhập môn Khoa học dữ liệu

Giảng viên: Đinh Xuân Trường
Học viện Công nghệ Bưu chính Viễn thông
Khoa Công nghệ thông tin 1

Hà Nội, 2024
http://www.ptit.edu.vn

Thông tin giảng viên

Giảng viên Bộ môn Khoa học máy tính -
Học viện Công nghệ Bưu chính Viễn thông - CNTT1

- Thạc sỹ Khoa học - Hệ thống thông tin PFIEV
- Nghiên cứu: Khoa học dữ liệu, học máy & đồ thị
- Giảng dạy: Kiến trúc máy tính, Hệ điều hành, Xây dựng hệ thống nhúng, Lập trình Python
- Liên hệ: truongdx@ptit.edu.vn

Nội dung

1 Sơ lược về Khoa học dữ liệu

2 Nội dung chính

3 Tài liệu tham khảo

4 Phương pháp đánh giá

5 Hướng dẫn về bài tập nhóm - phần lý thuyết

Sơ lược về Khoa học dữ liệu

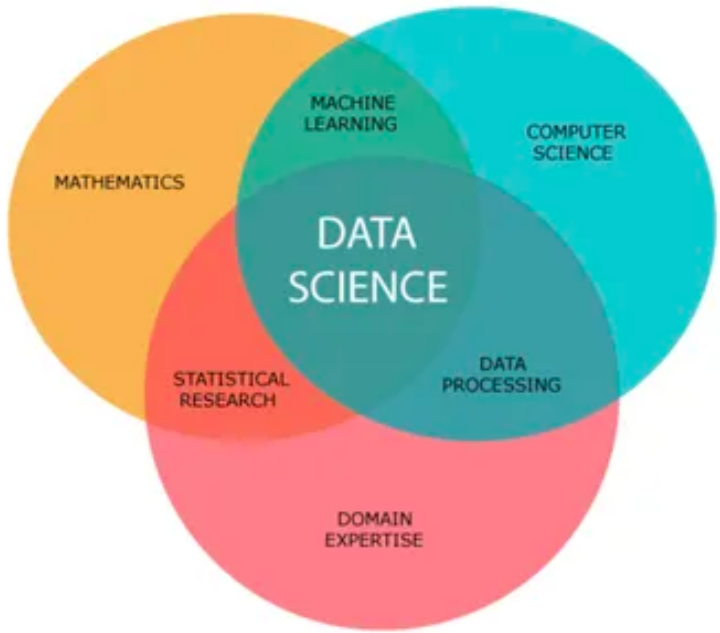
Sơ lược về Khoa học dữ liệu

Khoa học dữ liệu là một lĩnh vực liên ngành sử dụng các phương pháp, quy trình, thuật toán và hệ thống khoa học để trích xuất tri thức và hiểu biết từ các dữ liệu có cấu trúc và không có cấu trúc. Nói một cách đơn giản, khoa học dữ liệu là việc thập, xử lý, tổ chức và phân tích dữ liệu để có được hiểu biết từ dữ liệu cho nhiều mục đích khác nhau.

Sơ lược về Khoa học dữ liệu

Khoa học dữ liệu (Data Science) phụ thuộc vào mật thiết bởi 3 thành phần :

- Toán học (Đại số, giải tích, xác suất thống kề và tối ưu hoá)
- Công nghệ thông tin,
- Hiểu biết về ngành, lĩnh vực.



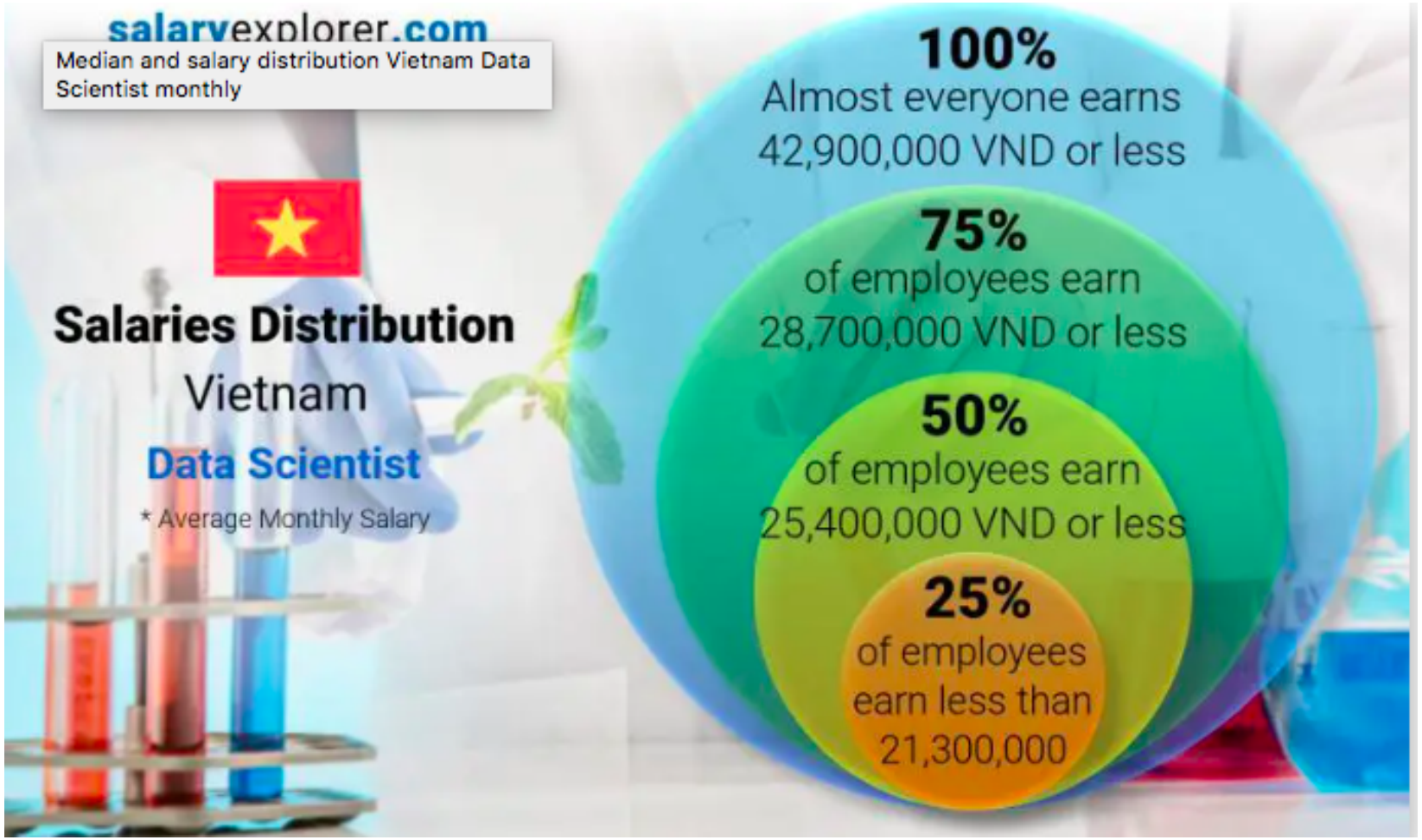
Các cấp độ của Khoa học dữ liệu

Khoa học dữ liệu là một lĩnh vực rộng nhưng được chia nhỏ thành nhiều phần nhỏ để giải quyết một vấn đề cụ thể, nó được chia thành 3 cấp độ:

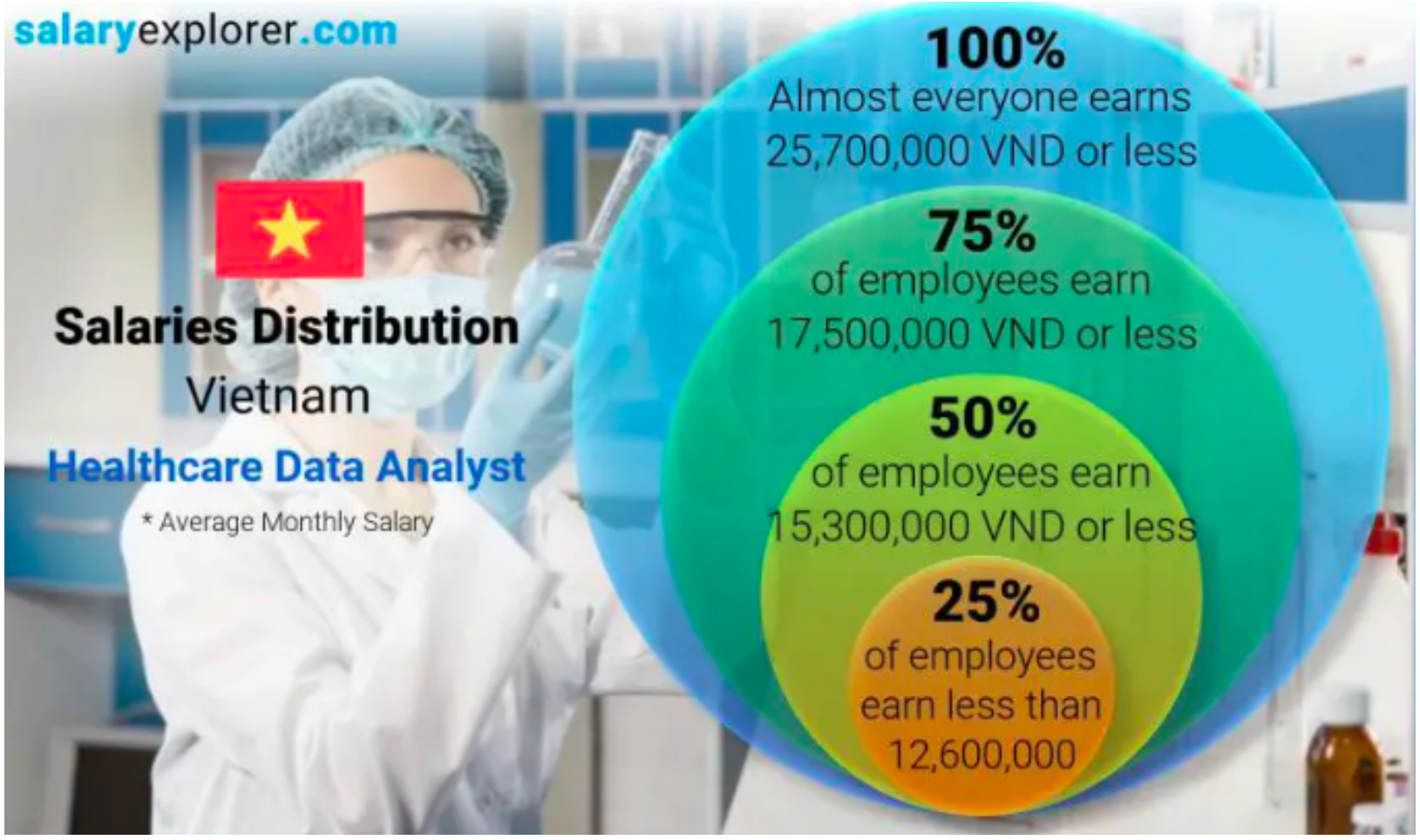
- Cấp độ 1: Tìm kiếm và định vị dữ liệu dựa trên các công nghệ và thuật toán thống kê.
- Cấp độ 2: Sử dụng công nghệ phân tích để chuyển đổi dữ liệu thành các nhóm thông tin cụ thể.
- Cấp độ 3: Hệ thống hoá và xây dựng cấu trúc cho thông tin cấp độ 2. Những thông tin này sẽ được đào tạo chuyên sâu về phân tích, kết nối và thể hiện thông tin có tính ứng dụng cao.

Làm thế nào để trở thành một Data Scientist

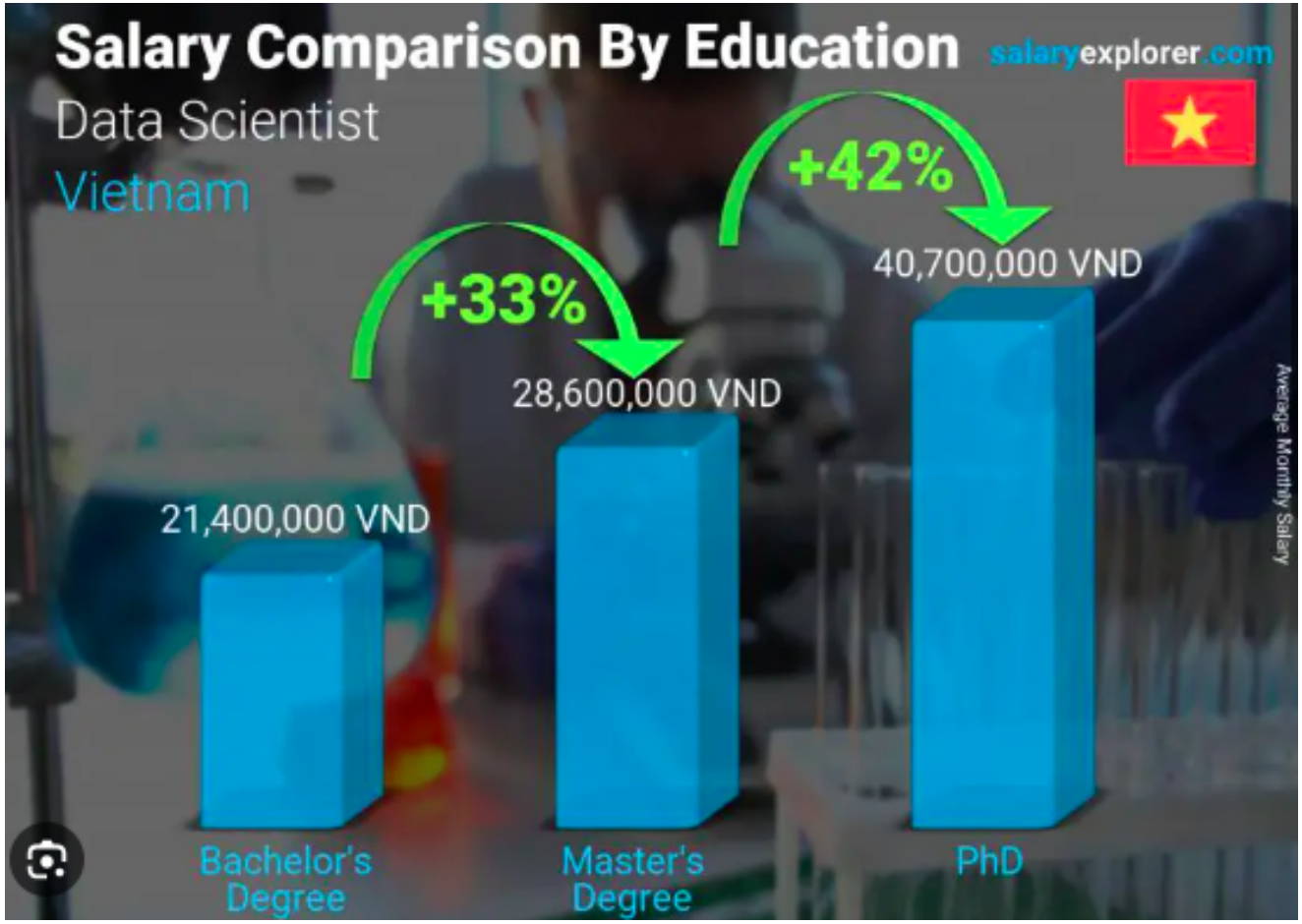
Tại sao bạn cần trở thành một data scientist?



So sánh với phân tích dữ liệu (data analyst)?



So sánh lương với bằng cấp Data Scientist



Nội dung

1 Sơ lược về Khoa học dữ liệu

2 Nội dung chính

3 Tài liệu tham khảo

4 Phương pháp đánh giá

5 Hướng dẫn về bài tập nhóm - phần lý thuyết

Chương 1: Kiến thức nền tảng

- **Đại số tuyến tính**
  - Vectors
  - Ma trận
  - Định thức ma trận
- **Lý thuyết xác suất**
  - Sự khiện và không gian mẫu
  - Công thức tính toán xác suất
  - Biến cố độc lập và thống kê Bayes
  - Biến ngẫu nhiên
  - Phân bố đều liên tục
  - Phân phối chuẩn
- **Lý thuyết thống kê**
  - Khoảng tin cậy
  - Kiểm định giả thuyết

Chương 2: Chuẩn bị dữ liệu

- **Thu thập và xử lý dữ liệu**
- **Làm sạch dữ liệu**
  - Xử lý các mục nhập bị thiếu
  - Xử lý các mục nhập không chính xác và không nhất quán
- **Co giãn và chuẩn hóa dữ liệu**
- **Giảm chiều và biến đổi dữ liệu**
  - Lấy mẫu
  - Lựa chọn đặc trưng
  - Giảm chiều dữ liệu

Chương 3: Trực quan hóa dữ liệu

- Đồ thị dạng đường thẳng
- Đồ thị điểm rời rạc
- Trực quan hóa lỗi
- Đồ thị đường viền
- Histograms và mật độ
- Văn bản và chú thích
- Đồ thị ba chiều
- Dữ liệu địa lý

Chương 4: Học máy

- **Các khái niệm cơ ản**
  - Học và suy diễn
  - Đánh giá mô hình
  - Bias vs Variance
  - Overfitting vs Underfitting
- **Biến đổi và trích chọn đặc trưng**
  - Classify Characteristics
  - Text Characteristics
  - Image characteristics
- **Các loại học máy**
  - Supervised learning
  - Unsupervised Learning
  - Semi-Supervised Learning
  - Reinforcement Learning

Chương 5: Cơ sở dữ liệu (SQL & noSQL)

- **Cơ sở dữ liệu cơ bản**
  - CREATE TABLE và INSERT
  - UPDATE
  - DELETE
  - SELECT
  - GROUP BY
  - ORDER BY
  - JOIN
- **Cơ sở dữ liệu nâng cao**
  - Subquery
  - Query optimization
  - NoSQL

Chương 6: Hệ khuyến nghị

- Giới thiệu
- Lọc dựa trên nội dung
- Lọc cộng tác
- Các hệ khuyến nghị lai
- Các hệ khuyến nghị dựa trên ngữ cảnh
- Khuyến nghị theo phiên

Nội dung

1 Sơ lược về Khoa học dữ liệu

2 Nội dung chính

3 Tài liệu tham khảo

4 Phương pháp đánh giá

5 Hướng dẫn về bài tập nhóm - phần lý thuyết

Tài liệu tham khảo

1. Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường *Bài giảng nhập môn Khoa học dữ liệu*, Khoa học máy tính - CNTT1 - PTIT, 2023.
2. Joel Grus, *Data Science from Scratch: First Principles with Python*, O'Reilly Media, 2nd edition, 2019.
3. Jake VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, 2017.
4. Charu C. Aggarwal, *Data Mining: The Textbook*, Springer International Publishing Switzerland, 2015.

Link Tài liệu tham khảo:

https://drive.g...

Nội dung

1 Sơ lược về Khoa học dữ liệu

2 Nội dung chính

3 Tài liệu tham khảo

4 Phương pháp đánh giá

5 Hướng dẫn về bài tập nhóm - phần lý thuyết

Phương pháp đánh giá

- **Điểm thành phần**
  - Chuyên cần 10%
  - Bài tập nhóm 20% (10% - 1 bài kiểm tra và 10% Project hằng tuần theo nhóm)
  - Kiểm tra giữa kỳ 10% (Bài tập nhóm - Lý thuyết)
  - Thi cuối kỳ 60%

Thiếu một điểm thành phần (bài tập, bài kiểm tra giữa kỳ), hoặc nghỉ quá 20% tổng số giờ của môn học, không được thi hết môn.

Nội dung

1 Sơ lược về Khoa học dữ liệu

2 Nội dung chính

3 Tài liệu tham khảo

4 Phương pháp đánh giá

5 Hướng dẫn về bài tập nhóm - phần lý thuyết

Hướng dẫn về bài tập nhóm - phần lý thuyết

Mỗi nhóm sẽ được chia thành viên ngẫu nhiên để trình bày các chủ đề dưới đây theo mỗi tuần.

- **Chủ đề 1:** *Ôn tập về Đại số tuyến tính, giải tích ma trận và tính toán xác suất.* Tài liệu tham khảo:
  - 1.*Machine Learning cơ bản*, Vũ Hữu Tiệp - **{Phần 1.1, 1.2, 1.3}**
  - 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 4, Chương 6}**
  - 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Phần 1.1 và 1.2}**
- **Chủ đề 2:** *Thống kê, kiểm định giả thuyết và suy luận: thống kê Maximum Likelihood và Maximum A Posteriori.* Tài liệu tham khảo:
  - 1. *Machine Learning cơ bản* - Vũ Hữu Tiệp - **{Phần 1-3, 1-4}**
  - 2. Nhập môn thống kê hướng tới học máy - Phạm Minh Hoàng

Hướng dẫn về bài tập nhóm - phần lý thuyết

- 3. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 5}**
- 4. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Phần 1.3}**

* **Chủ đề 3:** *Thu thập và chuẩn bị dữ liệu.* Tài liệu tham khảo:
  - 1. *Data Mining: The Textbook*, Springer International Publishing Switzerland, Charu C. Aggarwal - **{Chương 18}**
  - 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 9}**
  - 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Phần 2.1}**
* **Chủ đề 4:** *Tiền xử lý dữ liệu và trích chọn đặc trưng.* Tài liệu tham khảo:

Hướng dẫn về bài tập nhóm - phần lý thuyết

- 1. *Data Mining: The Textbook*, Springer International Publishing Switzerland, Charu C. Aggarwal - **{Chương 2}**
- 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 10}**
- 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Phần 2.2, 2.3, 3.4}**
- 4. *Machine Learning cơ bản* - Vũ Hữu Tiệp - **{Chương 6}**
- **Chủ đề 5:** *Giới thiệu về học máy: Phân loại, Hyperparameters và Model Validation, Overfitting - Underfitting, Bias - Variance* . Tài liệu tham khảo:
  - 1. *Data Mining: The Textbook*, Springer International Publishing Switzerland, Charu C. Aggarwal - **{Chương 8 và chương 9}**
  - 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 11}**

Hướng dẫn về bài tập nhóm - phần lý thuyết

- 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Phần 4.1}**
- 4. *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, Jake VanderPlas - **{Chương 5}**
- 5. *Machine Learning cơ bản* - Vũ Hữu Tiệp - **{Chương 8}**
- **Chủ đề 6:** *Các thuật toán ML về Regression: sử dụng công cụ Orange Data Mining* Tài liệu tham khảo:
  - 1. *Machine Learning cơ bản* - Vũ Hữu Tiệp - **{Chương 7, Chương 14, chương 15}**
  - 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 14, 15, 16, 12, 13, 17}**
  - 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Chương 4}**

Hướng dẫn về bài tập nhóm - phần lý thuyết

- 4. *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, Jake VanderPlas - **{Chương 5}**
- **Chủ đề 7:** *Các thuật toán Phân loại và phân cụm: sử dụng công cụ Orange Data Mining* Tài liệu tham khảo:
  - 1. *Machine Learning cơ bản* - Vũ Hữu Tiệp - **{Phần 3}**
  - 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{từ Chương 13 đến Chương 20}**
  - 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Chương 4}**
  - 4. *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, Jake VanderPlas - **{Chương 5}**
  - 5. *Data Mining: The Textbook*, Springer International Publishing Switzerland, Charu C. Aggarwal - **{Chương 10, 11}**

Hướng dẫn về bài tập nhóm - phần lý thuyết

- **Chủ đề 8:** *Neural Network* Tài liệu tham khảo:
  - 1. *Machine Learning cơ bản* - Vũ Hữu Tiệp - **{Phần 4}**
  - 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 19}**
  - 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Chương 4}**
  - 4. *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, Jake VanderPlas - **{Chương 5}**
- **Chủ đề 9:** *Xử lý dữ liệu văn bản (Text mining) hoặc dữ liệu time series hoặc dữ liệu đồ thị (graph data)/ phân tích dữ liệu mạng xã hội* Tài liệu tham khảo:
  - 1. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 21, 22}**
  - 2. *Data Mining: The Textbook*, Springer International Publishing Switzerland, Charu C. Aggarwal - **{Chương 13, 14, 15, 16, 17, 19 }**

Hướng dẫn về bài tập nhóm - phần lý thuyết

- **Chủ đề 10:** *Giảm chiều dữ liệu và tối ưu hoá hàm lồi* Tài liệu tham khảo:
  - 1. *Machine Learning cơ bản* - Vũ Hữu Tiệp - **{Phần 4,7,8}**
  - 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 19}**
  - 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Chương 4}**
  - 4. *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, Jake VanderPlas - **{Chương 5}**
- **Chủ đề 11:** *Dữ liệu lớn - Big Data và các thuật toán truy vấn trên dữ liệu lớn (Ranking & Search Engine* Tài liệu tham khảo:
  - 1. *Mining of Massive Datasets*, Jure Leskovec - **{Chương 2, Chương 3}**
  - 2. *Data Mining: The Textbook*, Springer International Publishing Switzerland, Charu C. Aggarwal - **{Chương 18 }**

Hướng dẫn về bài tập nhóm - phần lý thuyết

- 3. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 25}**
- **Chủ đề 12:** *Hệ khuyến nghị* Tài liệu tham khảo:
  - 1. *Machine Learning cơ bản* - Vũ Hữu Tiệp - **{Phần 5}**
  - 2. *Data Science from Scratch: First Principles with Python*,Joel Grus, O'Reilly Media - **{Chương 23}**
  - 3. *Bài giảng nhập môn Khoa học dữ liệu* Nguyễn Kiều Linh, Vũ Hoài Nam, Đinh Xuân Trường - **{Chương 6}**
  - 4. *Data Mining: The Textbook*, Springer International Publishing Switzerland, Charu C. Aggarwal - **{Chương 18}**

Ngoài các tài liệu tham khảo trên, sinh viên có thể tự tìm kiếm các tài liệu tương ứng để phục vụ cho mục đích tìm hiểu và triển khai các phần tính toán lý thuyết và lập trình.

H-1 Hướng dẫn Cách thức trình bày về bài tập nhóm - phần lý thuyết

Số lượng thành viên được chia mặc định trình bày hằng tuần theo thời gian 60 phút bao gồm các nội dung sau:

- Phần tổng hợp lý thuyết và bài tập minh hoạ (nếu có)
- Phần vận dụng (sử dụng các công cụ và lập trình sử dụng các framework để trực quan hoá phần lý thuyết)
- **Đầu ra cho hoạt động này bao gồm:**
  - Trình bày trên lớp (yêu cầu tất cả các thành viên của nhóm phải tham gia trình bày - **PHẢI** sử dụng slide template theo hướng dẫn trang 33 dưới đây).
  - Quyển báo cáo tổng hợp tất cả các nội dung đã tìm hiểu - **PHẢI VIẾT THEO MẪU SAU**: Mẫu quyển Báo cáo **KHÔNG** nhận file dưới bất kỳ hình thức khác.

H-2 Mẫu Slide trình bày phần lý thuyết

Sinh viên chọn một trong các mẫu dưới đây để sử dụng cho phần Slide trình bày:

- Mẫu 1: Màu đỏ tươi PTIT Presentation Red (Unofficial - RUG): Link Red template
- Mẫu 2: Màu đỏ rượu Merlot PTIT Presentation Red Merlot (Unofficial - RUG): : Link Red Merlot template
- Mẫu 3: Màu xanh Navy PTIT Presentation Navy (Unofficial - AIC): Link Navy Template
- Mẫu 4: Màu xanh đậm PTIT Presentation Midnight Blue (Unofficial - AIC): Link Midnight Blue Template

H-3 Các mốc thời gian nộp tài liệu - Phần trình bày

Mỗi nhóm sẽ có các mốc thời gian nộp các tài liệu như sau:

- **Về Slide trình bày** nộp về Link Google Driver của Nhóm lớp với các yêu cầu sau:
  - Tên của Slide đặt theo cú pháp: *Topic X - Nhóm Y - Lớp N0Z - Presentation.pdf*
  - Thời gian nộp: trước 23h59 sau ngày trình bày 1 ngày. Ví dụ Nhóm 1 trình bày vào sáng thứ 6 ngày 30/08/2024 thì cần nộp slide trước 23h59 thứ 7 ngày 31/08/2024.
- **Về Quyển báo cáo** nộp về Link Google Driver của Nhóm lớp với các yêu cầu sau:
  - Tên của Slide đặt theo cú pháp: *Topic X - Nhóm Y - Lớp N0Z - Document.pdf*
  - Thời gian nộp: trước 23h59 sau ngày trình bày 1 tuần. Ví dụ Nhóm 1 trình bày vào sáng thứ 6 ngày /08/2024 thì cần nộp slide trước 23h59 thứ 6 ngày 06/09/2024.

H-4 Hướng dẫn phần Project - Cập nhật hằng tuần

Các nhóm lựa thành viên mỗi nhóm tối đa 6 thành viên và chọn đề tài bao gồm việc thu thập dữ liệu trên internet để giải quyết một bài toán thực tế nào đó:

- Đề tài cần đưa ra vào tuần số 3 và được giảng viên chấp nhận đề tài (cần trao đổi để xác nhận đề tài có phù hợp không?)
- *Các nhóm cập nhật tiến độ theo hàng tuần bằng cách sử dụng template báo cáo sau :* Mẫu quyển, **KHÔNG** nhận file dưới bất kỳ hình thức khác.
- **Mỗi tuần các nhóm cần cập nhật kết quả của topic tuần đó trước buổi học của tuần tiếp theo.**
- Nhóm tự tạo link như mẫu trên và yêu cầu gửi link latex trên về email của giảng viên với tiêu đề email theo cú pháp *[KHDL_N0Y] Nhóm X Gửi link báo cáo hằng tuần*.