CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

Bài 03 - A

QUY TRÌNH XÂY DỰNG MÔ HÌNH MÁY HỌC

Machine learning Pipeline

TS. Nguyễn Vinh Tiệp

NỘI DUNG

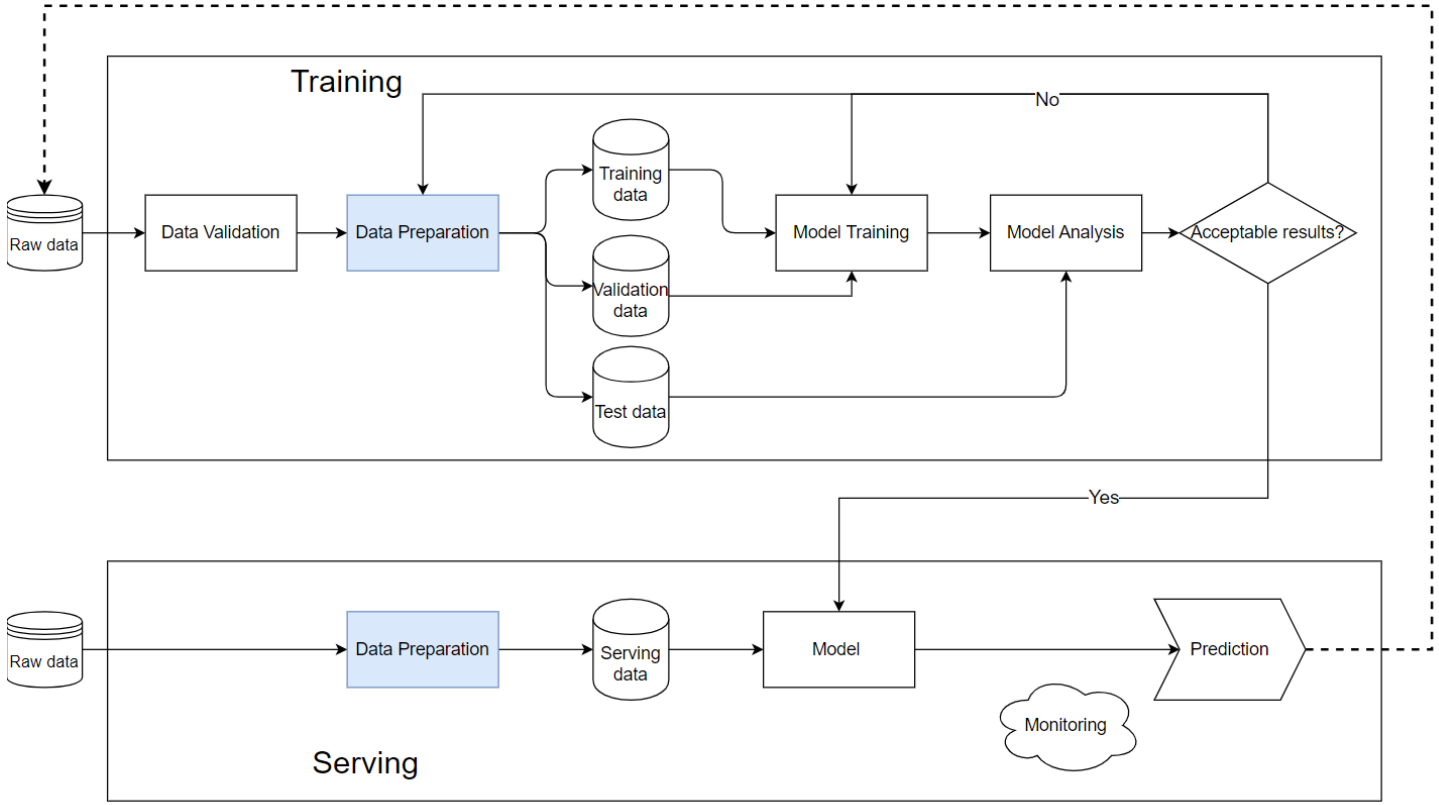
1. Machine Learning Pipeline ?
2. Exploratory Data Analysis (EDA)?

What-why-how


What
Why
How
What: Vấn đề/đối tượng là gì?
Why: Tại sao phải dùng đối tượng đó? Tại sao phải giải quyết vấn đề đó?
How: Chúng ta phải làm như thế nào?

ML Pipeline là gì?

- ML pipeline là một cách để mã hóa và tự động hóa quy trình làm việc của mô hình ML
- Bao gồm nhiều bước tuần tự thực hiện mọi thứ từ trích xuất dữ liệu, xử lý dữ liệu đến huấn luyện, đánh giá và triển khai mô hình



https://machinelearningcoban.com/tabml_book/ch_intro/pipeline.html

Data preparation

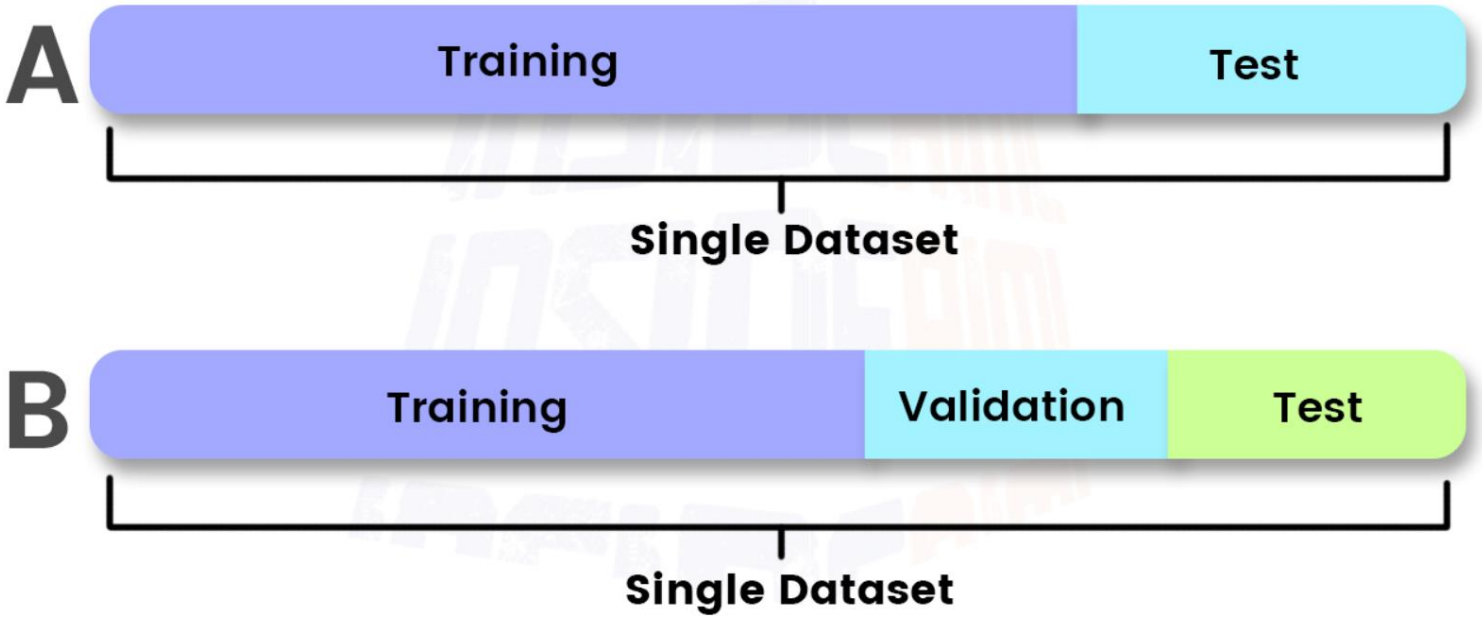
- Data fusion
- Data cleaning
- Data augmentation
- Data visualization
- Data splitting
- ...

Data splitting

- Training data → huấn luyện mô hình
- Testing data → kiểm tra hiệu suất mô hình
- Validation data → tối ưu hyper-parameters

Data splitting

- Ví dụ: train : test : valid = 7: 2 : 1

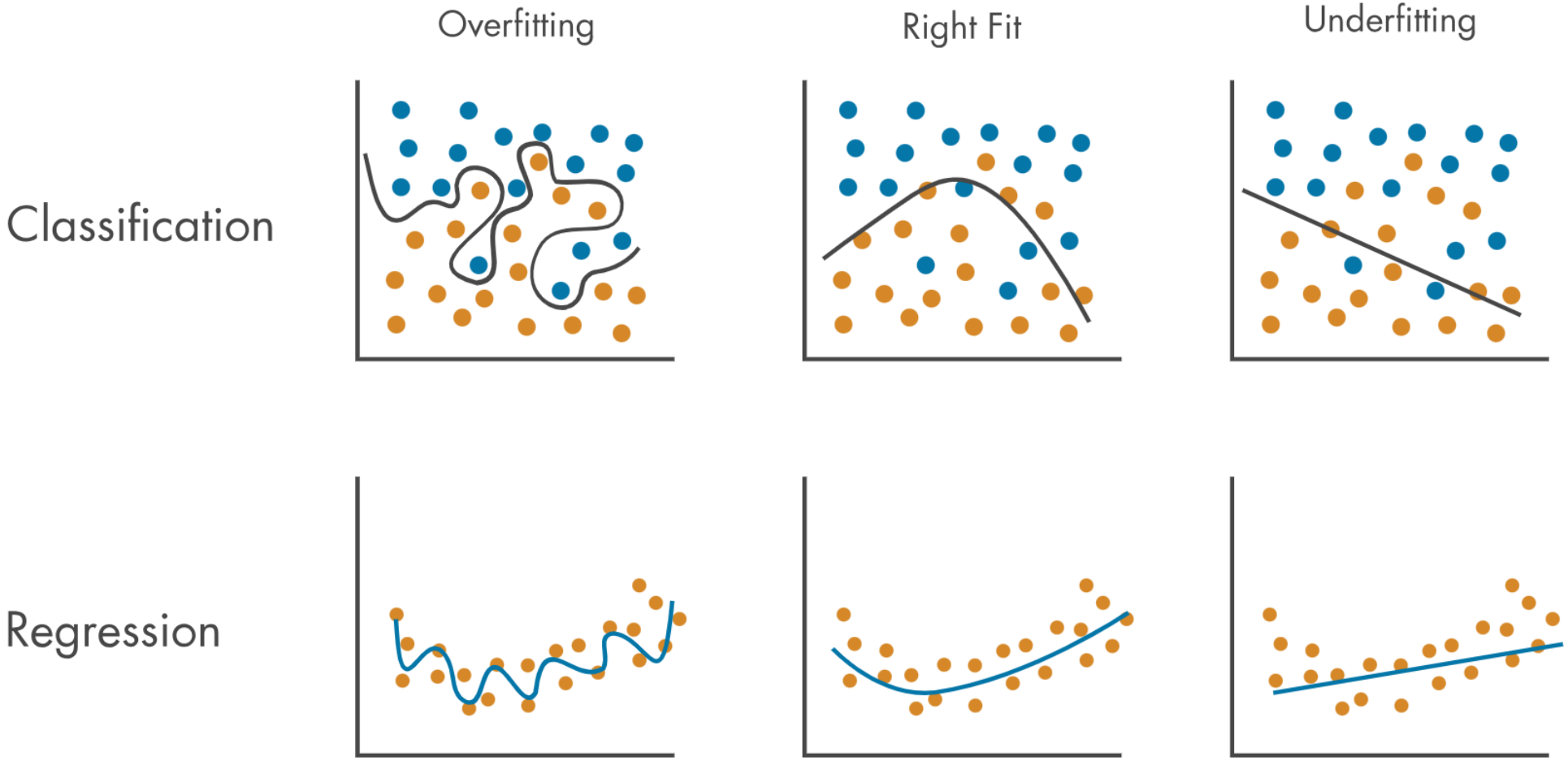


https://insideaiml.com/blog/Everything-you-need-to-know-about-Splitting-Dataset-1156

Model training

- Lựa chọn mô hình ML (bài toán)
- Phương pháp tối ưu mô hình (deep learning)
- Chọn hyper-parameter

Model overfiting



Model overfiting

Error	Overfitting	Right Fit	Underfitting
Training	Low	Low	High
Test	High	Low	High

Model analysis

- Đánh giá hiệu suất mô hình dựa trên số liệu và hình ảnh
- Tùy chỉnh hyper-parameter

Tại sao chọn ML Pipeline?

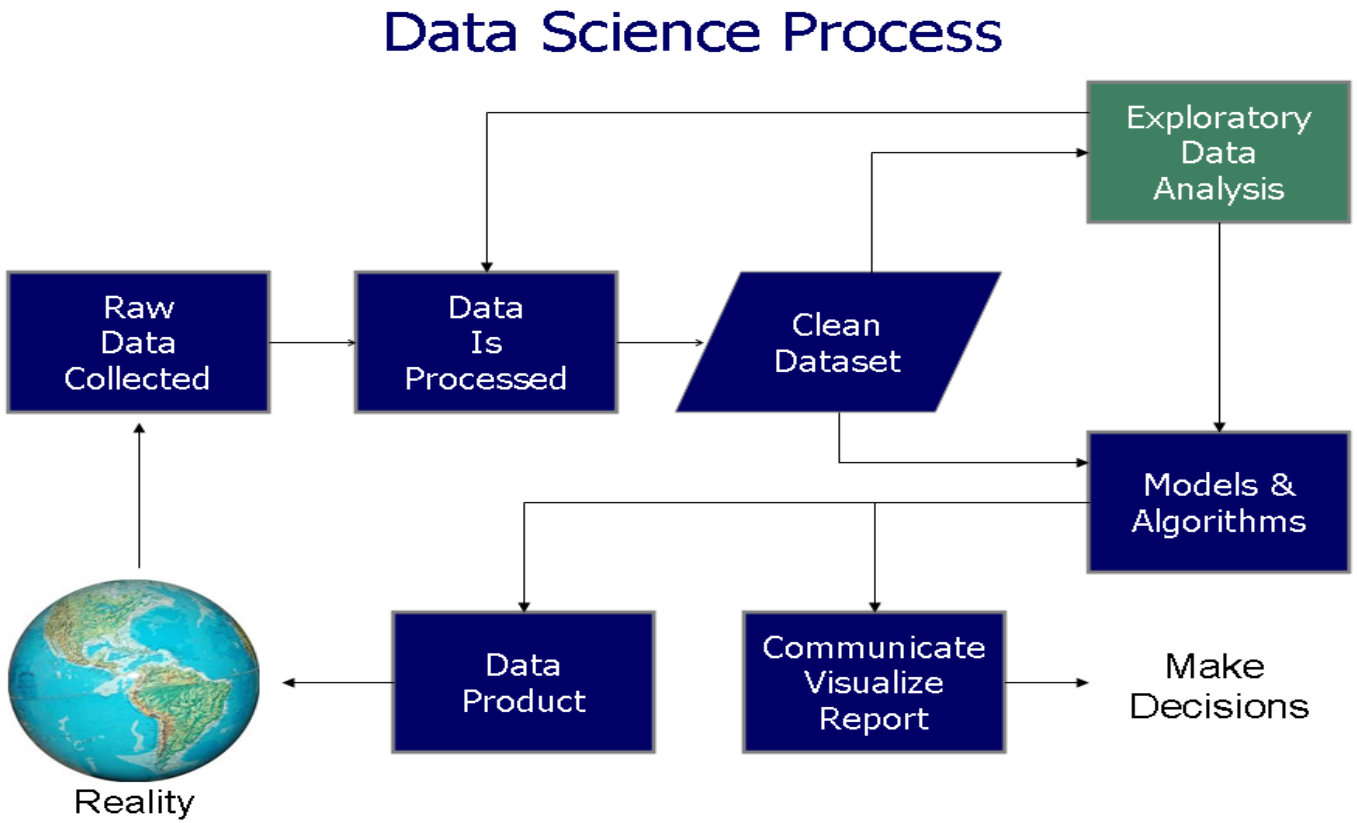
01	Tiêu chuẩn hóa và hiệu quả	Tự động hóa và chuẩn hóa quy trình xây dựng, đánh giá và triển khai các mô hình ML → tiết kiệm thời gian và nguồn lực
02	Khả năng lặp lại	Làm cho nó có thể tái tạo kết quả của một thử nghiệm học máy
03	Độ lặp lại	Làm cho nó có thể và đáng tin cậy để lặp lại cùng một quá trình nhiều lần
04	Hợp tác	Toàn bộ quá trình được mã hóa trong mã → chia sẻ ý tưởng và làm việc cùng nhau

Tại sao chọn ML Pipeline?

05	Khả năng tái sử dụng mã và tính mô-đun	Quản lý thời gian tốt hơn và ít lỗi hơn Nếu một phần của pipeline cần phải được thay đổi?
06	Kiểm soát và giám sát phiên bản	Cho phép các nhà phát triển theo dõi các thay đổi và đảm bảo rằng họ đang làm việc với phiên bản mới nhất
07	Đơn giản hóa việc triển khai sản xuất	Mô hình được đào tạo, tối ưu hóa và sẵn sàng triển khai vào sản xuất → tất cả các bước cần thiết đã sẵn sàng

Giới thiệu về EDA

- Trong thống kê, EDA là một cách tiếp cận phân tích các tập dữ liệu để tóm tắt các đặc điểm chính của chúng, thường sử dụng đồ họa thống kê và các phương pháp trực quan hóa dữ liệu khác



https://en.wikipedia.org/wiki/Exploratory_data_analysis

EDA - Phân tích dữ liệu thăm dò

Tại sao chúng ta cần EDA ?

- Hiểu dữ liệu
  Xác định các mẫu
  - Biết số lượng đặc trưng, loại dữ liệu đặc trưng, phân phối dữ liệu
  - Xác định các mẫu, mối quan hệ hoặc xu hướng trong dữ liệu có thể không rõ ràng ngay lập tức.
- Làm sạch dữ liệu
  - Xác định các vấn đề với dữ liệu: giá trị bị thiếu hoặc trùng lặp, ngoại lệ, lỗi hoặc sự không nhất quán
  - Quyết định cách xử lý những vấn đề này: xóa chúng hoặc điền dữ liệu bị thiếu hoặc giữ lại những vấn đề này

EDA - Phân tích dữ liệu thăm dò

- Tại sao chúng ta cần EDA ?
  - Kỹ thuật rút trích đặc trưng
    - Xác định các đặc trưng và tương tác quan trọng có thể hữu ích trong việc xây dựng các mô hình ML
    - Xác định các cơ hội để tạo ra các đặc trưng mới có thể cải thiện hiệu suất của các mô hình
  - Ngăn chặn kết quả gây hiểu lầm &; Truyền thông
    - EDA giúp ngăn chặn kết quả sai lệch bằng cách làm quen với dữ liệu trước
    - Trực quan hóa dữ liệu theo những cách khác nhau, có thể giúp truyền đạt những phát hiện của bạn cho người khác, đặc biệt là cho các bên liên quan phi kỹ thuật.

Mục tiêu của EDA

Mục tiêu của EDA là:

- Cho phép khám phá bất ngờ trong dữ liệu
- Đề xuất các giả thuyết về nguyên nhân của các hiện tượng quan sát được
- Đánh giá các giả định dựa trên suy luận thống kê nào
- Hỗ trợ lựa chọn các công cụ và kỹ thuật thống kê phù hợp
- Cung cấp cơ sở để thu thập thêm dữ liệu thông qua khảo sát hoặc thí nghiệm

QUIZ & CÂU HỎI