Phương pháp luận trong KHDL

Date: @March 18, 2024, **Writer: @Khôi Nguyễn**

Link Notion: PPLKHDL Notion

Recall

Phương pháp luận

- Who ? (của ai ?)
- What ? (giúp làm gì ?)
- How ? (thực hiện như thế nào ?)

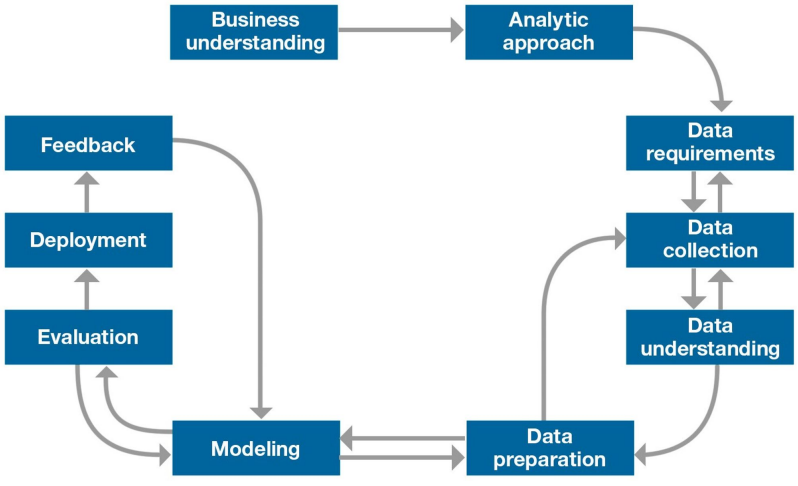
Phần 1:

- What ? (bao nhiêu giai đoạn? là gì ?)
- How ? (giải quyết những gì ?)

Notes

▼ *Phương pháp luận*

- Một phương pháp của John Rollins được sử dụng trong khoa học dữ liệu.
- Xây dựng thành một luồng công việc, giúp không bị phân vân khi quyết định nên làm gì tiếp theo và đảm bảo việc đang làm là đúng đắn.
- Với 3 phần, gồm 10 câu hỏi ứng với 10 giai đoạn:



▼ ***Phần 1: Xác định hướng giải quyết***

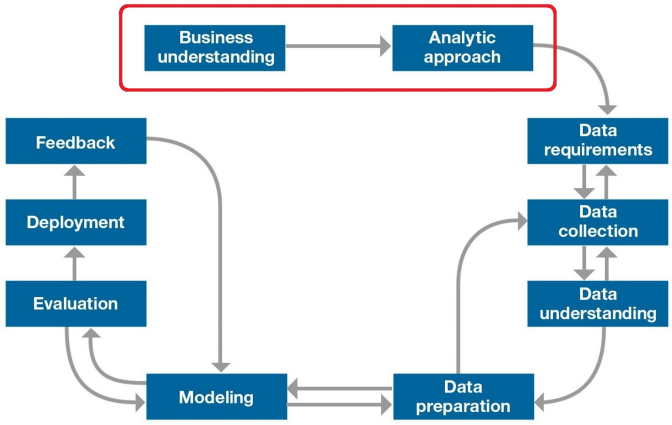
Phần 2:

- Who ? (Xử lý với cái gì ?)
- What ? (bao nhiêu giai đoạn? là gì ?)
- How ? (giải quyết những gì ?)

Phần 3:

- Who ? (Xử lý với cái gì ?)
- What ? (bao nhiêu giai đoạn? là gì ?)
- How ? (giải quyết những gì ?)

Gồm 2 giai đoạn:



▼ **Business understanding**

1. Vấn đề gặp phải là gì ?
   - Giai đoạn quan trọng, định hình cả dự án.
   - Làm rõ vấn đề cần giải quyết, xác định dữ liệu cần thiết.

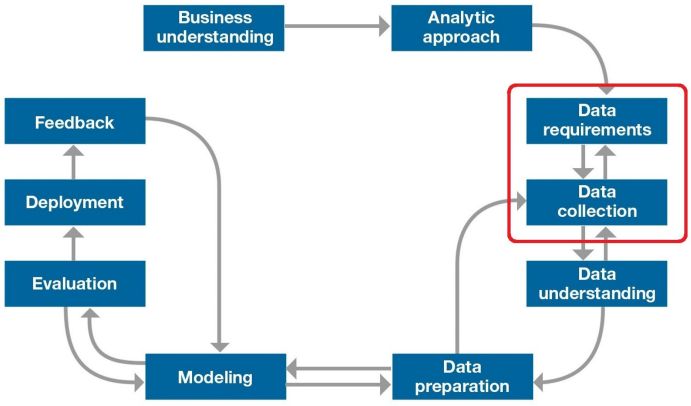
▼ **Analytic approach**

2. Sử dụng dữ liệu như thế nào để trả lời câu hỏi ?
   - Giải quyết quyết vấn đề đã được xác định bên trên.
   - Xác định mô hình cần thiết để giải quyết vấn đề hiệu quả nhất.

▼ ***Phần 2: Xử lý dữ liệu***

Gồm 4 giai đoạn:

Giai đoạn làm việc với dữ liệu



▼ Data requirements

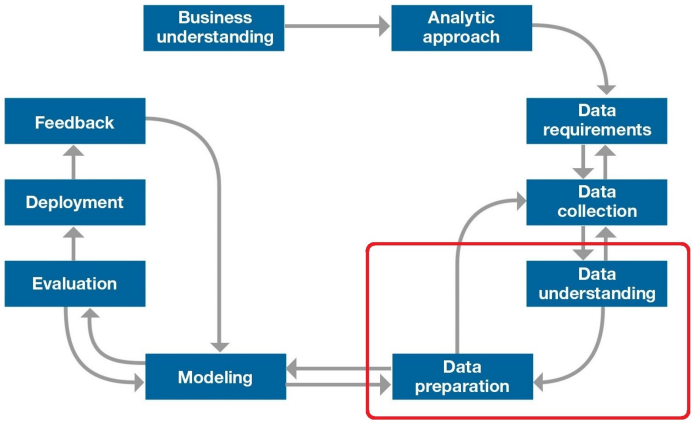
3. Cần dữ liệu nào ?

- Xác định dữ liệu cần thiết để giải quyết vấn đề.
- Dữ liệu là chìa khoá tìm ra giải pháp.
- Xác định dữ liệu bao gồm:
  - Xác định nội dung
  - Xác định định dạng
  - Xác định nguồn gốc

▼ Data collection

4. Dữ liệu từ đâu ? Lấy bằng cách nào ?

- Xác định yêu cầu và thu thập dữ liệu.
- Đánh giá và điều chỉnh dữ liệu thu thập.
- Sử dụng phương pháp đánh giá và trực quan hoá.
- Xác định và xử lý lỗ hổng dữ liệu.



▼ Data understanding

5. Dữ liệu đó có giải quyết được vấn đề không ?

- Xây dựng tập dữ liệu và kiểm tra tính đại diện cho vấn đề cần giải quyết.
- Các phương pháp thường dùng:
  - Thống kê mô tả
  - Tương quan theo cặp
  - Biểu đồ
- Chuẩn hoá dữ liệu.

▼ Data preparation

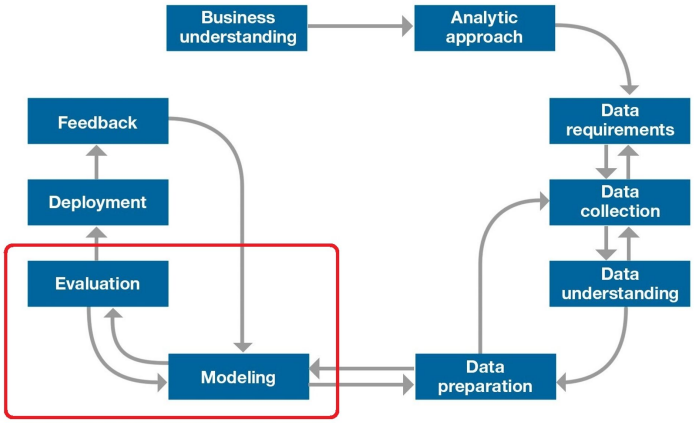
6. Cần làm gì thêm không ?

- Loại bỏ các đặc trưng dữ liệu không cần thiết.
- **Tự động hóa**: Quy trình thu thập và chuẩn bị dữ liệu có thể được tự động, giúp giảm thời gian và tập trung cho mô hình.

- **Transforming data**: để dễ dàng thao tác hơn (xử lý
  giá trị bị thiếu, không hợp lệ và trùng lặp).
- **Feature engineering**: sử dụng kiến thức về dữ liệu
  để tạo ra các tính năng giúp thuật toán học máy hoạt động hiệu quả.
- **Phân tích văn bản**: Cần được thực hiện để mã hóa dữ
  liệu văn bản và tạo ra các nhóm thích hợp.

▼ ***Phần 3: Triển khai mô hình***

Gồm 4 giai đoạn:



▼ Modeling

7. Mô hình hoá dữ liệu như thế nào để giải quyết vấn đề ?

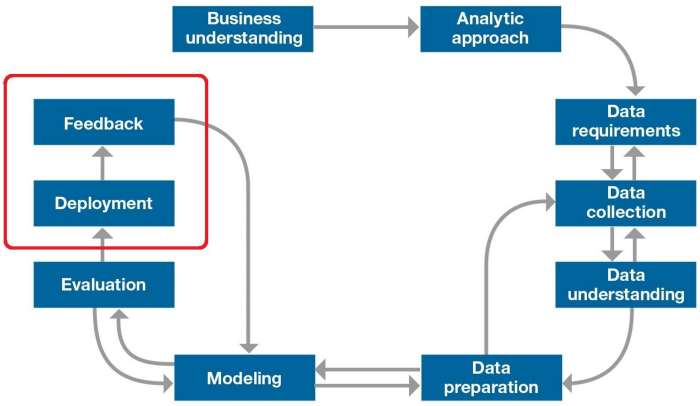
- Phát triển các mô hình mô tả hoặc dự đoán dựa trên phân tích đã thực hiện.

- **Tranning set**: Tập dữ liệu với kết quả đã biết
  trước, dùng để đánh giá và điều chỉnh mô hình.
- Quá trình mô hình hóa đòi hỏi hiểu biết sâu sắc về vấn đề, phương
  pháp phân tích phù hợp, và liên tục cải tiến và điều chỉnh mô hình để
  đạt được kết quả chắc chắn.

▼ Evaluation

7. Mô hình có giải quyết được vấn đề không ? Cần chỉnh gì không ?

- Được thực hiện liên tục trong quá trình xây dựng mô hình.
- Đánh giá chất lượng mô hình và có đáp ứng yêu cầu ban đầu
  không.
- 2 giai đoạn:
  - Thực hiện biện pháp chuẩn đoán.
  - Kiểm tra ý nghĩa thống kê.



▼ Deployment

8. Có thể triển khai mô hình vào thực tiễn không ?

- Cung cấp một cách thức dễ dàng để các bên liên quan sử dụng thông qua các ứng dụng web hoặc tích hợp vào phần mềm.

▼ Feedback

9. Những phản hồi mang tính xây dựng để giải quyết vấn đề là gì ?

- Trôi dạt dữ liệu là hiện tượng dữ liệu thay đổi theo thời gian, ảnh hưởng đến hiệu suất dự đoán của mô hình.
- Phản hồi của người dùng giúp đánh giá mô hình được triển khai có hoạt động tốt không.
- Phản hồi giúp điều chỉnh mô hình để vẫn đáp ứng được nhu cầu của người dùng.

▼ — — SUMMARY — —

- Phương pháp luận trong Khoa học dữ liệu của John Rollins.
- Gồm 3 phần (Vấn đề, dữ liệu, mô hình) với 10 giai đoạn.
- Giúp việc triển khai mô hình từ đầu đến cuối diễn ra trơn tru và giải quyết được vấn đề đã ban đầu.