Khai thác dữ liệu

1. Khai thác dữ liệu và phân tích dự đoán là gì

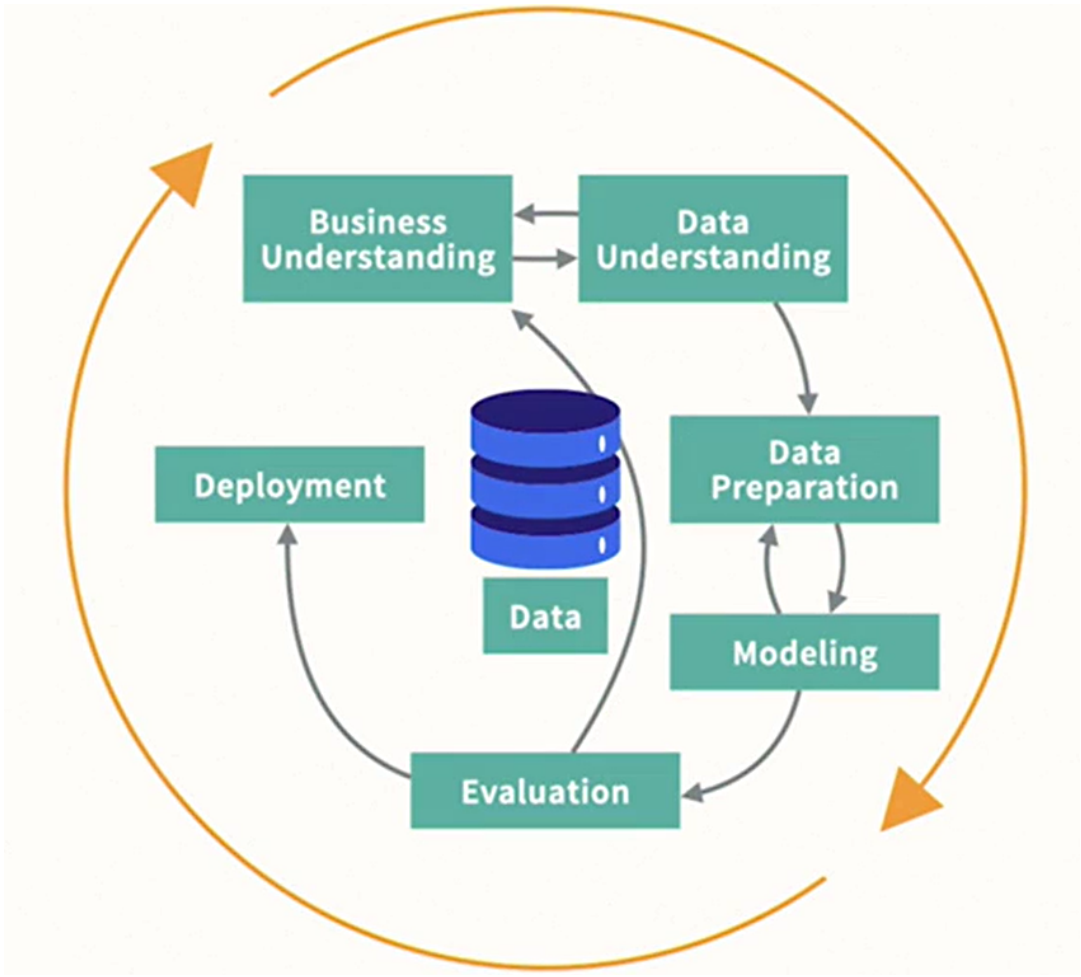
   ▼ Giới thiệu các yếu tố cần thiết

   ▼ Xác định khai thác dữ liệu

   - Khai thác dữ liệu là tìm ra các mẫu trong dữ liệu quá khứ và sau đó tận dụng các mẫu đó trên dữ liệu hiện tại để đưa ra những dữ đoán trong tương lai
   - Khai thác dữ liệu là việc lựa chọn và phân tích dữ liệu tích lũy trong quá trình kinh doanh bình thường, để tìm và xác nhận các mối quan hệ chưa biết trước đây có thể tạo ra kết quả tích cực và có thể kiểm chứng được thông qua việc triển khai các mô hình dự đoán khi áp dụng cho dữ liệu mới

   ▼ Giới thiệu CRISP-DM

   - CRISP-DM là một từ viết tắt của Cross- Industry Standard Process For Data Mining
   - Hướng dẫn chi tiết về khai thác dữ liệu



2. Định nghĩa vấn đề

▼ Bắt đầu với bước đầu tiên vững chắc: Định nghĩa vấn đề

- Khi các nhà phân tích cố gắng xây dựng các mô hình dự đoán họ thường đưa ra một yêu cần nghiên cứu một cách thiếu cân nhắc. Định nghĩa vấn đề kém gần như chắc chắn là lý do lớn nhất khiến các dự án thất bại

▼ Đóng khung vấn đề về một quyết định vi mô

- Những quyết định vi mô nghĩa là khi đang đưa ra một quyết định rất cụ thể về một trường hợp duy nhất.

▼ Tại sao mọi mô hình đều cần một chiến lược can thiệp hiệu quả

- Chiến lược can thiệp là hành đồng thực hiện cho điểm cao mà không lấy điểm thấp

- ▼ Đánh giá tiềm năng của dự án bằng các chỉ số kinh doanh và ROI
- ▼ Chuyển các vấn đề kinh doanh thành các vấn đề khai thác dữ liệu

3. Yêu cầu dữ liệu
   - ▼ Hiểu các yêu cầu về dữ liệu
     - Ngăn ngừa những sai lầm
     - Cho phép đánh giá chính xác hơn
   - ▼ Thu thập dữ liệu quá khứ
   - ▼ Đáp ứng yêu cầu tệp phẳng
     - Các thuật toán học máy cần dữ liệu ở một hình thức cụ thể, vì vậy yêu cầu dữ liệu là một tập tin phẳng
   - ▼ Xác định biến mục tiêu
     - Xác định biến mục tiêu cũng giống như việc kết quả cuối cùng đã được xác định
   - ▼ Lựa chọn dữ liệu có liên quan
   - ▼ Gợi ý về tích hợp dữ liệu
     - Tích hợp dữ liệu (data integration) là một cái gì đó mà mọi dữ án cần
   - ▼ Hiểu kỹ thuật tính năng
   - ▼ Phát triển thủ công
4. Nguồn lực cần thiết
   - ▼ Bộ kỹ năng và tài nguyên sẽ cần
     - Kĩ năng làm việc nhóm rất cần thiết
     - Các tài nguyên
       - Các thuật toán khai thác dữ liệu chuyên biệt
       - Một nhóm đa chức năng
       - Nhiều thời gian trong một thời gian biểu thực tế
       - Tiếp cận các chuyên gia về chủ đề

- ▼ So sánh máy học và thống kê
    - Máy học mô hình được thiết kế để đưa ra dự đoán chính xác nhất có thể. Các thống kê mô hình được thiết kế để suy luận về mối quan hệ giữa các biến
- ▼ Đánh giá yêu cầu của nhóm
    - Lập một nhóm với các thành viên có nền tảng và chuyên môn khác nhau tạo trên sự đa dạng trong công việc
- ▼ Lập ngân sách đủ thời gian
- ▼ Làm việc với các chuyên gia về chủ đề

5. Những vấn đề sẽ đối mặt
    - ▼ Dự đoán những thách thức của dự án
        - Dữ liệu bị thiếu
        - Sự chê bai của cấp trên
        - Các mô hình sẽ bị xuống cấp theo thời gian
    - ▼ Giải quyết dữ liệu bị thiếu
    - ▼ Giải quyết sự chê bai của cấp trên
    - ▼ Giải quyết các mô hình xuống cấp
6. Tìm giải pháp
    - ▼ Chuẩn bị cho các nhiệm vụ giai đoạn mô hình hóa
    - ▼ Tìm kiếm các giải pháp tối ưu
    - ▼ Tìm kiếm kết quả bất ngờ
        - Quá tiết kiệm với công cụ dự đoán, loại bỏ biến này, loại bỏ biến kia. Từ đó làm giảm thông tin chi tiết mà các thao tác không lường trước có thể được thực hiện
    - ▼ Thiết lập bằng chứng rằng mô hình hoạt động
        - Cần thử hoạt động mô hình
    - ▼ Áp dụng phương pháp thử nghiệm và lỗi sai

7. Đưa ra giải pháp để làm việc
    - ▼ Chuẩn bị cho giai đoạn triển khai
    - ▼ Sử dụng xác suất và xu hướng
    - ▼ Hiểu mô hình meta
    - ▼ Hiểu khả năng tái sử dụng
    - ▼ Chuẩn bị triển khai mô hình
    - ▼ Cách tiếp cận tài liệu dữ án
8. Chín quy luật khai thác dữ liệu
    - ▼ CRISP-DM và luật khai thác dữ liệu
    - ▼ Hiểu CRISP-DM
        - Chia thành 6 giai đoạn
            - Hiểu biết kinh doanh
            - Hiểu biết dữ liệu
            - Chuẩn bị dữ liệu
            - Mô hình hóa
            - Đánh giá
            - Phát triển
    - ▼ Lời khuyên khi sử dụng CRISP-DM
        - Hiểu biết kinh doanh
            - Thiết lập sự đồng thuận giữa các thành viên trong nhóm
        - Hiểu biết về dữ liệu
            - Giai đoạn này cần phải thực hiển vì nó rất quan trọng. Còn là giai đoạn ở giữa giai đoạn hiểu biết kinh doanh và chuẩn bị dữ liệu. Công việc này kiểm tra khả năng của dữ liệu để giải quyết vấn đề kinh doanh và nếu dữ liệu thiếu ở đâu, tạo một danh sách cải thiện cần làm và những cái tiến cần thực hiện trong giai đoạn chuẩn bị dữ liệu
        - Chuẩn bị dữ liệu

        - Giải quyết danh sách cái công việc cần làm trong giai đoạn hiểu biết dữ liệu đặt ra
    - Mô hình hóa
        - Đừng bỏ qua 3 giai đoạn đầu tiên
    - Đánh giá
        - Không phải là nhiệm vụ đánh giá mô hình mà là đánh giá kinh doanh
        - Nên sử dụng ngôn ngữ của doanh nghiệp
    - Phát triển

▼ Hiểu chín luật khai thác dữ liệu

▼ Hiểu luật thứ nhất và thứ hai

- Luật 1
    - Mục tiêu kinh doanh là nguồn gốc của mọi giải pháp khai thác dữ liệu
- Luật 2
    - Giải quyết mọi sự hiểu lầm tiềm ẩn

▼ Hiểu luật chuẩn bị dữ liệu

▼ Hiểu luật về các mẫu

▼ Hiểu luật dự đoán và hiểu biết sâu sắc

▼ Hiểu quy luật giá trị

▼ Hiểu lý do tại sao các mô hình thay đổi