

Mô tả mô hình CRISP-DM:

CRISP-DM (viết tắt của Cross-Industry Standard Process for Data Mining) là một mô hình quy trình chuẩn, được sử dụng rộng rãi trong các dự án khai phá dữ liệu. Mục tiêu của CRISP-DM là cung cấp một khuôn khổ có cấu trúc, linh hoạt và lặp lại để hướng dẫn các chuyên gia phân tích dữ liệu từ khi bắt đầu dự án cho đến khi triển khai kết quả cuối cùng.

Mô hình CRISP-DM bao gồm sáu giai đoạn chính và không phải là một quy trình tuyến tính cứng nhắc. Các giai đoạn thường được lặp lại để tinh chỉnh kết quả.

Mô hình CRISP-DM gồm sáu giai đoạn sau:

- **Giai đoạn 1: Hiểu biết về bài toán (Business Understanding)**

Đây là giai đoạn đầu tiên và quan trọng nhất, tập trung vào việc hiểu rõ mục tiêu và yêu cầu của dự án từ góc độ kinh doanh. Xác định mục tiêu kinh doanh và mục tiêu khai phá dữ liệu. Đánh giá tình hình, nguồn lực và rủi ro sau đó lập kế hoạch tổng thể cho dự án.

- **Giai đoạn 2: Hiểu biết về dữ liệu (Data Understanding)**

Giai đoạn này tập trung vào việc thu thập, khám phá và xác minh chất lượng dữ liệu. Thu thập dữ liệu ban đầu từ nhiều nguồn khác nhau. Mô tả, kiểm tra và khám phá dữ liệu để tìm ra các mẫu và điểm bất thường. Xác minh chất lượng dữ liệu để tìm ra các giá trị thiếu hoặc không chính xác.

- **Giai đoạn 3: Tiền xử lý dữ liệu (Data Preparation)**

Đây là giai đoạn tốn thời gian nhất, nhằm làm sạch và chuẩn bị dữ liệu để sẵn sàng cho việc mô hình hóa. Nó lựa chọn dữ liệu và làm sạch các giá trị thiếu hoặc bất thường. Xây dựng các thuộc tính mới và tích hợp dữ liệu từ nhiều nguồn và định dạng dữ liệu để phù hợp với các thuật toán.

- **Giai đoạn 4: Xây dựng mô hình (Modeling)**

Giai đoạn này là lúc áp dụng các thuật toán khai phá dữ liệu để xây dựng mô hình. Giai đoạn này ta sẽ chọn kỹ thuật mô hình hóa phù hợp với mục tiêu. Sau đó, huấn luyện mô hình với dữ liệu đã chuẩn bị và đánh giá hiệu suất ban đầu của mô hình.

- **Giai đoạn 5: Đánh giá (Evaluation)**

Giai đoạn này tập trung vào việc đánh giá kết quả một cách toàn diện và xác định xem mô hình có thực sự giải quyết được vấn đề kinh doanh ban đầu hay không. Phân tích kết quả của mô hình và so sánh với các mô hình khác nhau. Và xem xét lại toàn bộ quy trình để đưa ra quyết định có nên triển khai mô hình hay không.

- **Giai đoạn 6: Triển khai (Deployment)**

Đây là giai đoạn cuối cùng, đưa mô hình đã hoàn thiện vào sử dụng trong môi trường thực tế. Lên kế hoạch sẵn để tích hợp mô hình vào hệ thống. Thiết lập các kế hoạch giám sát kỹ càng và bảo trì mô hình. Tạo báo cáo tổng kết và trình bày kết quả cuối cùng.