Quy trình CRISP-DM (Cross Industry Standard Process for Data Mining) là một quy trình chuẩn, linh hoạt, được ứng dụng rộng rãi trong các dự án Khai phá dữ liệu nói riêng và Khoa học dữ liệu nói chung. Quy trình có khung tham chiếu và hướng dẫn thực hành có sẵn nhiều tài liệu. Dieu này giúp cho các nhà phân tích triển khai dự án có hệ thống, linh hoạt và lặp lại được. Quy trình này gồm sáu giai đoạn chính, tạo thành chu trình khép kín từ bước xác định bài toán đến các bước triển khai.

Giai đoạn đầu đầu tiên là Business Understanding( Hiểu biết nghiệp vụ), bước mà các nhà phân tích sẽ tập trung xác định mục tiêu kinh doanh, hiểu rõ bối cảnh, vấn để của bài toán thực tế. Đây là bước quan trọng vì nó phải mang lại giá trị thực tiễn cho tổ chức.

Tiếp theo là Data Understanding (Hiểu biết dữ lieu), giai đoạn này sẽ tập trung vào thu thập dữ liệu thô và tiến hành các bước khám phá để hiểu dữ liệu. Giai đoạn này sữ dụng các kỵ thuật thống kê mô tả, trực quan hóa và phân tích dữ liệu được áp dụng nhằm đánh giá chất lượng dữ liệu. Qua đó biết được thiếu sót dữ liệu, trùng lặp dữ liệu và các giá trị ngoại lệ. Giai đoạn này là tiền đề để phát hiện những vấn đề tiềm ẩn cần xử lý trong giai đoạn tiếp theo.

Giai đoạn thứ ba là Data Preparetion (Chuẩn bị dữ liệu), được coi là giai đoạn chiếm thòi gian trong một dự án. Giai đoạn này thường chiếm 70% thời gian để các nhà phân tích xử lý các giả trị thiếu, ngoại lệ, biến đổi dữ liệu, chuẩn hóa dữ liệu, mã hóa dữ liệu hay lựa chọn đặc trưng sẽ áp dụng để phù hợp với mô hình sẽ dủng. Cuối giai đoạn này, ta thu được một tập dữ liệu đã được xử lý cho huấn luyện và kiểm tra sẵn sáng cho việc xây dựng mô hình.

Sau khi dữ liệu đã được xử lý, các nhà phân tích sẽ tiến hành Modelling ( Xây dựng mô hình). Ở bước này, việc chọn 1 thuật toán phù hợp là yếu tố quyết định đễ thu được tri thức từ dữ liệu. Các thuật toán có thể dùng bao gồm phân loại (classification), hồi quy (regression), phân cụm (clustering) hoặc khai phá luật kết hợp (association rule mining). Mô hình được huấn luyện, tinh chỉnh tham số (hyperparameter tuning) và so sánh nhiều phương án để chọn ra mô hình tốt nhất.

Bước tiếp theo là Evaluation (Đánh giá mô hình), nơi mô hình được đánh giá hiệu suất dựa trên các chỉ số thường dùng như Accuracy, Precision, Recall, F1-score, RMSE, v.v Kết quả phân tích được so sánh với kỳ vọng thực tế để xác định mô hình có đủ tin cậy để triển khai hay không. Nếu chưa đạt yêu cầu, nhóm có thể quay lại các giai đoạn trước để điều chỉnh. Cuối cùng là Deloyment( Triển khai), giai đoạn đưa mô hình vào ứng dụng thực tế. Việc triển khai có thể dưới dạn báo cáo kết quả. Dashboard trực quan hoặc tích hợp mô hình vào hệ thống của doanh nghiệp. Hơn nữa, ta có thể theo dõi bảo trì và chỉnh sửa mô hình nhằm mang lại giá trị bền vững.