**Bài tập 1**

Đề bài: Mô tả mô hình CRISP-DM và các giai đoạn của qui trình, trong 1 trang word.

Bài làm

Mô hình CRISP-DM hay còn gọi là **Cross Industry Standard Process for Data Mining**, là một **khung quy trình tiêu chuẩn trong khai thác dữ liệu (data mining)**, được thiết kế để giúp các dự án phân tích dữ liệu đi theo một phương pháp luận có hệ thống, từ việc hiểu vấn đề kinh doanh đến triển khai giải pháp là một trong những chuẩn được sử dụng rộng rãi trong ngành khoa học dữ liệu. Bao gồm 6 giai đoạn, linh hoạt và có thể tinh chỉnh

**1. Business Understanding (Hiểu biết về vấn đề kinh doanh):**

* + Mục tiêu là hiểu rõ mục tiêu và vấn đề mà doanh nghiệp muốn giải quyết.
  + Các task gồm: Determine Business Objectives – xác định mục tiêu chính từ góc nhìn doanh nghiệp, ví dụ Primary goal: Inventory Optimization – tối ưu quản lý kho; Assess Situation – đánh giá tình hình hiện tại, các yếu tố ảnh hưởng, nguồn lực, hạn chế; Determine Data Mining Goals – chuyển đổi mục tiêu kinh doanh thành data mining goals, ví dụ Demand Forecasting, Market Basket Analysis, Customer Segmentation; Produce Project Plan – lập kế hoạch dự án.

**2. Data Understanding (Hiểu biết dữ liệu**):

* + Mục tiêu là khám phá dữ liệu và đánh giá chất lượng dữ liệu.
  + Các task gồm: Collect Initial Data – thu thập dữ liệu từ nhiều nguồn, ví dụ sales data, customer data, product data; Describe Data – mô tả dữ liệu: kiểu dữ liệu, phạm vi, số lượng, các trường thông tin; Explore Data – khám phá dữ liệu để tìm pattern, xu hướng, phân phối, ví dụ phân tích xu hướng bán trái cây theo mùa, phân bố độ tuổi khách hàng; Verify Data Quality – kiểm tra dữ liệu: missing values, duplicates, outliers, inconsistencies.

**3. Data Preparation (Chuẩn bị dữ liệu):**

* + Mục tiêu là chuẩn bị dữ liệu cuối cùng cho mô hình.
  + Các task gồm: Data Selection – lựa chọn các biến phù hợp; Data Cleaning – làm sạch dữ liệu, loại bỏ trùng lặp, xử lý missing values; Data Construction – tạo biến mới (feature engineering); Integrate Data – kết hợp dữ liệu từ nhiều nguồn thành dataset thống nhất; Format Data – chuẩn hóa dữ liệu, scaling, encoding, phân chia train/test set.

**4. Modeling (Mô hình hóa):**

* + Mục tiêu là xây dựng mô hình dự đoán, phân loại hoặc phân nhóm dữ liệu.
  + Các task gồm: Select Modelling Technique – chọn thuật toán phù hợp; Generate Test Design – lập kế hoạch đánh giá mô hình, ví dụ chia train/test, cross-validation; Build Model – huấn luyện mô hình, chọn tham số, chạy thuật toán; Assess Model – đánh giá và giải thích mô hình theo mục tiêu kinh doanh và tiêu chí thành công, ví dụ Time Series Forecasting với ARIMA, k-means để phân khúc khách hàng.

**5. Evaluation (Đánh giá):**

* Mục tiêu là xác định mô hình nào phù hợp và quyết định bước tiếp theo.
* Các task gồm: Evaluate Results – đánh giá kết quả mô hình trên test set, ví dụ accuracy, mean error rate, confusion matrix; Review Process – xem xét lại toàn bộ quá trình từ business understanding đến modeling; Determine Next Steps – quyết định triển khai, cải thiện dữ liệu/mô hình hoặc thử thuật toán khác.

**6. Deployment (Triển khai):**

* Mục tiêu là áp dụng mô hình vào thực tế và sử dụng insights từ data mining để cải thiện kinh doanh.
* Các task gồm: Plan Deployment – lên kế hoạch triển khai mô hình (dashboard, hệ thống tự động, báo cáo); Plan Monitoring and Maintenance – theo dõi hiệu quả mô hình, cập nhật khi cần; Produce Final Report – báo cáo kết quả, phương pháp và insights rút ra; Review Project – đánh giá tổng thể dự án, học hỏi kinh nghiệm cho các dự án sau. CRISP-DM là một quy trình lặp lại, tương tác, từ việc hiểu vấn đề kinh doanh đến triển khai mô hình và đưa ra quyết định dựa trên dữ liệu.