

Bài tập - Hãy cho biết khai thác dữ liệu là gì và cho tối thiểu 03

ví dụ thực tế có ứng dụng khai thác dữ liệu

Khai thác dữ liệu (Data Mining) là một lĩnh vực xuất hiện khi lượng dữ liệu ngày càng nhiều và phức tạp. Hiểu một cách gần gũi, đây là quá trình tìm kiếm những thông tin quan trọng ẩn bên trong các tập dữ liệu lớn mà mắt thường hoặc cách phân tích thông thường khó nhận ra. Người ta áp dụng nhiều kỹ thuật khác nhau, từ thống kê, học máy cho đến các phương pháp phân tích truyền thống, nhằm phát hiện ra các xu hướng, mối liên hệ hay mô hình có ích cho việc nghiên cứu hoặc ra quyết định.

Ví dụ thực tế có ứng dụng khai thác dữ liệu

Trong sinh học, đặc biệt là nghiên cứu gen, khai thác dữ liệu được dùng rất nhiều. Các nhà khoa học thu thập dữ liệu biểu hiện gen của nhiều loại mô hoặc nhiều trạng thái bệnh khác nhau. Vì số lượng gen rất lớn nên họ phải sử dụng kỹ thuật gom cụm để xem những gen nào có cách hoạt động giống nhau. Những nhóm gen được tìm thấy thường liên quan đến cùng một chức năng sinh học hoặc liên quan đến một loại bệnh nhất định. Đây là bước quan trọng giúp hiểu rõ hơn cơ chế bệnh lý và hỗ trợ nghiên cứu thuốc.

Trong lĩnh vực môi trường, dữ liệu khí hậu từ các trạm đo trên khắp thế giới được lưu lại liên tục trong nhiều năm. Những dữ liệu này gồm nhiệt độ, lượng mưa, áp suất... và rất khó để nhìn ra quy luật nếu chỉ xem bằng cách thủ công. Vì vậy, người ta sử dụng các kỹ thuật như luật kết hợp hoặc phân loại để tìm ra xu hướng biến đổi khí hậu hoặc dự đoán các hiện tượng thời tiết bất thường. Các mô hình này giúp việc dự báo chính xác hơn và hỗ trợ xây dựng các chính sách thích ứng với biến đổi khí hậu.

Một ví dụ khác quen thuộc hơn là trong lĩnh vực tài chính, cụ thể là phát hiện gian lận giao dịch thẻ tín dụng. Ngân hàng lưu lại toàn bộ lịch sử giao dịch của khách hàng và dựa vào đó để xây dựng mô hình phân loại nhằm nhận biết những giao dịch bất thường. Khi hệ thống học được cách phân biệt giao dịch hợp lệ và nghi ngờ gian lận, nó có thể cảnh báo sớm để giảm thiệt hại cho cả ngân hàng lẫn người dùng.

Khai thác dữ liệu cũng được ứng dụng rất mạnh trong thương mại điện tử. Các sàn mua sắm hay nền tảng xem phim đều lưu lại hành vi của người dùng như sản phẩm đã xem, đã mua hoặc các nội dung đã tương tác. Từ những thông tin này, hệ thống sử dụng kỹ thuật lọc cộng tác để đo mức độ giống nhau giữa người dùng hoặc giữa các mặt hàng. Nhờ đó, mỗi người sẽ nhận được những gợi ý phù hợp hơn, ví dụ như “sản phẩm tương tự”, “người khác cũng mua”... Điều này vừa giúp doanh nghiệp tăng doanh số, vừa cải thiện trải nghiệm người dùng.

Như vậy, dù ứng dụng trong sinh học, môi trường, tài chính hay thương mại điện tử, mục tiêu chung của khai thác dữ liệu vẫn là tìm ra điều có ích từ những dữ liệu tương tự như vô tổ chức. Nhờ các phương pháp như gom cụm, phân loại hay luật kết hợp, dữ liệu thô được chuyển thành thông tin có giá trị, phục vụ thiết thực cho cả nghiên cứu lẫn đời sống.