BÀI 2 GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU

Học xong bài này, em sẽ:

- Nêu được sơ lược về mục tiêu và một số thành tựu của khoa học dữ liệu, nêu được ví dụ minh hoạ.

Có ý kiến cho rằng: Dữ liệu là tài sản quan trọng của tổ chức, doanh nghiệp. Theo em, nói như vậy là vì nguyên nhân nào sau đây:

1) Chi phí cao để thu thập, lưu trữ, bảo đảm an toàn dữ liệu.

2) Dữ liệu được sử dụng để tăng hiệu quả làm việc, tăng sức cạnh tranh của doanh nghiệp, tăng lợi ích kinh doanh.

1 Khoa học dữ liệu

Khi tìm kiếm với cụm từ khoá "Các lĩnh vực nghiên cứu của khoa học dữ liệu" thì có thể nhận được kết quả gồm có: Khai phá dữ liệu, Thống kê, Học máy, Phân tích,... Hãy nêu tên một lĩnh vực mà em hiểu biết nhiều nhất và cho biết lĩnh vực này nghiên cứu gì.

a) Khái niệm khoa học dữ liệu

Thuật ngữ "Khoa học dữ liệu" ban đầu phát sinh trong môi trường kinh doanh thông minh BI (Business Intelligence). Khoa học dữ liệu là bước phát triển tiếp theo của khoa học thống kê, khai phá dữ liệu, phát hiện tri thức trong dữ liệu,...

Khoa học dữ liệu là lĩnh vực liên ngành sử dụng các phương pháp khoa học, quy trình, công cụ của các ngành như Toán học và thống kê, Khoa học máy tính kết hợp với kiến thức chuyên môn trong các lĩnh vực ứng dụng như kinh doanh, tài chính ngân hàng, y tế, giáo dục,... nhằm rút ra được những hiểu biết sâu sắc từ dữ liệu (*Hình 1*). Khoa học dữ liệu là khoa học về việc quản trị và phân tích dữ liệu, trích xuất các giá trị, phát hiện tri thức từ dữ liệu phục vụ mục đích ra quyết định và lập kế hoạch. Khoa học dữ liệu giúp tăng hiệu quả, tăng cơ hội thành công, giảm rủi ro thất bại trong các hoạt động của tổ chức doanh nghiệp.

Hình 1. Minh hoạ Khoa học dữ liệu là lĩnh vực liên ngành

Các mục tiêu cụ thể của khoa học dữ liệu gồm:

1. *Phân tích và trực quan hoá dữ liệu*: xem xét các mẫu, xu hướng trong tập dữ liệu để hiểu dữ liệu và biểu diễn dữ liệu một cách trực quan; giúp người dùng có cái nhìn tổng quan và nhận biết được những yếu tố quan trọng, từ đó phát hiện vấn đề cần giải quyết.

2. *Xây dựng mô hình dự đoán, dự báo*: Sử dụng dữ liệu để xây dựng mô hình có khả năng dự đoán sự kiện tương lai như: sự thay đổi doanh số, xuất hiện rủi ro, biến động về khách hàng,...

3. *Tối ưu hoá quyết định*: điều chỉnh quyết định dựa trên dữ liệu, sử dụng các thuật toán tối ưu hoá để đưa ra quyết định tốt nhất.

4. *Phát hiện tri thức*: tìm ra các mối quan hệ, quy luật ẩn trong dữ liệu, xác định rõ nguyên nhân và kết quả, phát triển kiến thức mới.

b) Các giai đoạn của một dự án khoa học dữ liệu

Một dự án khoa học dữ liệu liên quan đến những vấn đề cụ thể mà tổ chức, doanh nghiệp cần giải quyết. Dự án được chia thành một số giai đoạn sau (*Hình 2*):

Hình 2. Các giai đoạn của một dự án khoa học dữ liệu

1. *Xác định vấn đề*: Hiểu rõ những vấn đề mà tổ chức, doanh nghiệp cần giải quyết. Từ đó, có thể xác định một số giả thuyết cần kiểm tra, đánh giá và quyết định.

2. *Thu thập dữ liệu*: Sau khi hiểu rõ vấn đề, cần thu thập dữ liệu liên quan từ nhiều nguồn khác nhau. Trong nhiều trường hợp, tập dữ liệu thu thập được thường rất lớn.

3. *Chuẩn bị dữ liệu*: Lựa chọn dữ liệu; tích hợp dữ liệu từ nhiều nguồn; làm sạch dữ liệu, xử lí các giá trị còn thiếu, không chính xác, loại bỏ ngoại lệ; biểu diễn dữ liệu dưới dạng phù hợp để sử dụng trong các mô hình phân tích.

4. *Phân tích và khai phá dữ liệu*: Áp dụng mô hình trên dữ liệu đã chuẩn bị để chọn lọc một số yếu tố quan trọng nhằm giải quyết vấn đề. Phân tích và khai phá dữ liệu nhằm tìm ra các mối quan hệ, quy luật ẩn trong dữ liệu để xây dựng các mô hình dự báo và phát triển kiến thức mới trong lĩnh vực hoạt động của tổ chức, doanh nghiệp.

5. *Đánh giá và giải thích*: Sử dụng các tiêu chí cụ thể để đánh giá chất lượng mô hình. Giải thích tác động của mô hình đến hoạt động của tổ chức, doanh nghiệp. Kiểm tra, đánh giá mô hình để triển khai.

6. *Ra quyết định và triển khai*: Sau các đánh giá nghiêm ngặt, kết quả phân tích

dữ liệu được trình bày cho cấp lãnh đạo quản lí tổ chức, doanh nghiệp để làm cơ sở ra quyết định và triển khai thực tế.

Một ví dụ minh hoạ

Lãnh đạo một cảng hàng không nhận thấy số lần máy bay chậm giờ cất cánh có xu hướng tăng là một vấn đề cần giải quyết. Một tổ dự án được giao nhiệm vụ đề xuất phương án cải tiến quy trình nghiệp vụ để giải quyết vấn đề trên. Một nhiệm vụ trong dự án là phân tích dữ liệu nhằm mục đích lập kế hoạch tốt hơn. Việc thực hiện nhiệm vụ này có thể coi là một dự án khoa học dữ liệu nhỏ.

Qua tìm hiểu thông tin sơ bộ, tổ dự án nhận thấy số lần máy bay chậm giờ cất cánh phụ thuộc vào số lượng hành khách qua sân bay và số lượng hành khách qua sân bay biến động tuỳ theo những khoảng thời gian khác nhau. Do đó, dự án cần phân tích dữ liệu để dự báo lượng hành khách qua sân bay trong tương lai. Từ đó, lập kế hoạch ngắn hạn giúp phân bổ tốt hơn nguồn lực đáp ứng yêu cầu công việc có biến động theo thời gian. Đây là một ví dụ về xác định vấn đề.

Tổ dự án cần thu thập các số liệu thống kê liên quan để có thể giải quyết vấn đề đã xác định, ví dụ: số lượng hành khách qua sân bay theo từng thời điểm, số lần máy bay chậm giờ cất cánh và nguyên nhân. Đây là ví dụ về thu thập dữ liệu.

Chuẩn bị dữ liệu ở đây là xác định những thuộc tính đặc trưng nào cần được phân tích; xử lí các số liệu còn thiếu hay xoá bỏ các số liệu trùng lặp, không chính xác; biểu diễn dữ liệu dưới dạng phù hợp để sẵn sàng áp dụng mô hình phân tích dữ liệu. Với ví dụ này, mô hình dữ liệu phù hợp là chuỗi thời gian.

Phân tích và khai phá dữ liệu trong trường hợp này là phân tích dự báo dùng chuỗi thời gian để dự báo số lượng hành khách qua sân bay trong các tháng tiếp theo.

Mô hình dự báo dữ liệu bằng chuỗi thời gian có tham số để xác định độ tin cậy của kết quả dự báo. Kết hợp độ tin cậy của mô hình lí thuyết và yêu cầu ứng dụng thực tế sẽ đánh giá được sự phù hợp để sử dụng trong việc lập kế hoạch ngắn hạn.

Đánh giá và giải thích tác động của việc cải tiến, đổi mới quy trình nghiệp vụ để thuyết phục lãnh đạo của tổ chức, doanh nghiệp ra quyết định triển khai là vấn đề có phạm vi rộng hơn, đòi hỏi phân tích dữ liệu theo một số khía cạnh khác.

2 Một số thành tựu của khoa học dữ liệu

Khoa học dữ liệu có nhiều ứng dụng trong kinh tế – xã hội.

Trong tài chính ngân hàng, khoa học dữ liệu giúp đánh giá rủi ro, phát hiện gian lận, lập mô hình đầu tư, phân khúc khách hàng.

Trong chăm sóc sức khoẻ, khoa học dữ liệu giúp dự đoán dịch bệnh, cải thiện chất lượng chăm sóc bệnh nhân, quản lí dịch vụ y tế, chế tạo thuốc chữa bệnh.

Trong sản xuất kinh doanh, khoa học dữ liệu giúp đưa ra các quyết định tầm chiến lược, tối ưu hoá quy trình để sản xuất kinh doanh, cá nhân hoá trải nghiệm của khách hàng và đưa ra khuyến nghị cho khách hàng. Trong dịch vụ công nghệ thông tin, khoa học dữ liệu giúp tối ưu hoá hệ thống thông tin, đảm bảo an ninh mạng,...

Các hệ thống trí tuệ nhân tạo như trợ lí ảo được phát triển đều có phần đóng góp của những dự án khoa học dữ liệu. Những tập dữ liệu lớn được thu thập, phân tích để hiểu rõ các thách thức, xây dựng các mô hình và huấn luyện đạt hiệu quả cho phép sử dụng trong thực tế.

Khoa học dữ liệu đạt được một số thành tựu đáng chú ý như sau đây.

a) Dự án Bộ gen người HGP

Dự án Bộ gen người HGP (Human Genome Project) kéo dài 13 năm (từ 1990 đến 2003) và tiêu tốn khoảng 3 tỉ USD là một nỗ lực quốc tế lớn nhằm nghiên cứu cấu trúc và chức năng của các gen trong bộ gen người. Dự án giúp xác định các biến thể di truyền, tạo nền tảng xác định mối quan hệ giữa các đột biến và đặc điểm sinh học. Lập bản đồ gen và giải trình tự gen là hai kĩ thuật để nghiên cứu cấu trúc và chức năng của gen. Kết quả của dự án đã mở ra một kỉ nguyên mới cho lĩnh vực khoa học sức khoẻ. Bộ gen người được tạo thành từ khoảng 3 tỉ cặp base. Giải trình tự một bộ gen thường sinh ra khoảng một trăm gigabyte dữ liệu. Giải trình tự nhiều bộ gen người có thể sinh ra hàng trăm petabyte dữ liệu. Để phân tích dữ liệu hệ gen người, cần phát triển các thuật toán có tốc độ nhanh và sử dụng các máy tính mạnh. Phân tích dữ liệu hệ gen người giúp các nhà nghiên cứu hiểu rõ hơn về cách thức hoạt động của gen, chức năng của gen, mối quan hệ giữa gen và đặc điểm sinh học, sức khoẻ, bệnh tật.

Trong đại dịch Covid-19, các nhà nghiên cứu đã sử dụng máy giải trình tự gen tiên tiến để nhanh chóng xác định virus SARS-CoV-2 ngay từ đầu đại dịch. Điều này đã cho phép phân tích và hiểu rõ hơn về cách virus gây ra bệnh, từ đó phát triển các phương pháp chẩn đoán, điều trị và phòng ngừa hiệu quả.

b) Các dự án nghiên cứu và khám phá không gian vũ trụ

Loài người không ngừng nghiên cứu và khám phá không gian vũ trụ với mục tiêu tìm kiếm các hành tinh có tiềm năng duy trì sự sống. Sau đây là một số dự án của NASA, cơ quan của chính phủ Mỹ về nghiên cứu hàng không và thám hiểm không gian vũ trụ.

Kính thiên văn Kepler, trong 9 năm hoạt động, đã tạo ra khoảng 678 GB dữ liệu, ghi lại độ sáng của khoảng 150 nghìn ngôi sao.

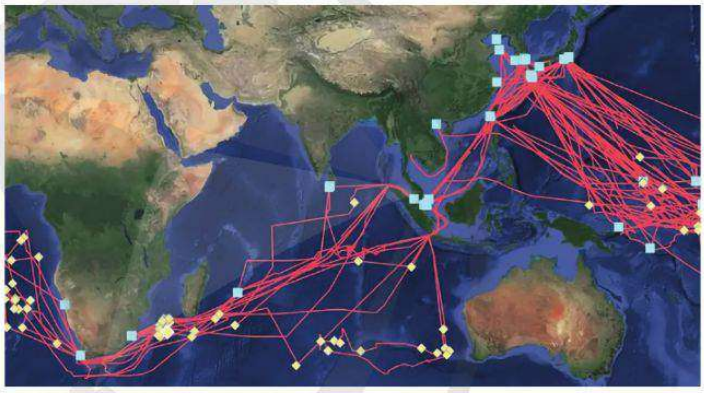
Các vệ tinh như TESS, K2, Plato thu thập các thông tin về hành tinh như khối lượng, kích thước, mật độ và quỹ đạo, tạo ra khoảng 100 GB dữ liệu mỗi ngày.

Để phân tích lượng lớn dữ liệu như vậy, cần phát triển các thuật toán học máy chuyên biệt và phức tạp. Những thuật toán này giúp phân loại các đặc trưng của hành tinh, phát hiện thay đổi bất thường trong ánh sáng ngôi sao và suy luận về các hành tinh khác trong hệ các ngôi sao dựa trên thay đổi quỹ đạo.

Đến nay đã phát hiện được khoảng 3 200 hệ hành tinh quay quanh các ngôi sao trong tổng số khoảng 200 tỉ ngôi sao thuộc dải Ngân Hà và có khoảng 63 hành tinh được xác định có khả năng nuôi dưỡng sự sống.

c) Hệ thống Giám sát đánh bắt cá toàn cầu

Hệ thống Giám sát đánh bắt cá toàn cầu (Global Fishing Watch) của Google sử dụng dữ liệu vệ tinh để cung cấp thông tin cho việc ngăn chặn đánh bắt cá bất hợp pháp (*Hình 3*). Mỗi ngày, hàng triệu vị trí của các con tàu trên các tuyến đường thuỷ khắp thế giới được ghi lại, cho phép xác định mục đích chuyến đi của mỗi con tàu kèm với điểm xuất phát của nó từ quốc gia nào. Từ đó cho biết nơi đang diễn ra hoạt động đánh bắt cá theo thời gian thực để có thể xác định tàu nào đánh bắt cá bất hợp pháp và vào thời điểm cụ thể nào.



Hình 3. Theo dõi lịch sử hoạt động của tàu biển trên website của Globol Fishing Watch (Nguồn: https://globalfishingwatch.org)

d) Các mô hình ngôn ngữ lớn

Các mô hình ngôn ngữ lớn LLM (Large Language Models) là một loại mô hình AI được thiết kế đặc biệt để hiểu ngôn ngữ tự nhiên. Một trong những LLM nổi tiếng nhất là GPT-3, có 175 tỉ tham số. Số lượng tham số càng lớn, mô hình càng có thể hiểu và xử lí ngôn ngữ một cách tinh vi hơn. GPT đã đạt được thành tựu ấn tượng có tính cách mạng trong xử lí ngôn ngữ tự nhiên. Được đào tạo dựa trên lượng dữ liệu văn bản rất lớn, GPT có thể tạo ra người máy thông minh sánh ngang hoặc có thể vượt con người trong một số nhiệm vụ phức tạp.

e) Mô hình phát hiện gian lận của American Express

Dịch vụ thẻ tín dụng American Express đã đạt được thành công đáng kể trong việc phát hiện gian lận nhờ có khoa học dữ liệu. Năm 2014, lần đầu tiên American Express triển khai mô hình học máy để phát hiện gian lận đã giúp cải thiện 30% so với các hệ thống cũ. Năm 2017, American Express đã phát triển một công cụ xác thực nâng cao sử dụng sinh trắc học để xác định ai đang thực hiện giao dịch thẻ tín dụng. Công cụ này đã giúp giảm được 60% giao dịch gian lận. Theo báo cáo Nilson tháng 2 năm 2021, American Express đã duy trì tỉ lệ gian lận thấp nhất ở Mỹ trong 14 năm liên tiếp.

Câu 1. Khoa học dữ liệu có những mục tiêu cụ thể gì?

Câu 2. Dự án khoa học dữ liệu gồm những giai đoạn nào?

Câu 3. Hãy nêu ví dụ về sự đóng góp của khoa học dữ liệu vào một thành tựu khoa học công nghệ.

Theo em, khoa học dữ liệu có thể đóng góp cho cải tiến quản lí giao thông đô thị để giảm tắc đường hay không? Giải thích.

Hãy cho biết mỗi phát biểu sau đây về khoa học dữ liệu là đúng hay sai:

a) Khoa học dữ liệu nhằm rút ra được những hiểu biết sâu sắc từ dữ liệu.

b) Khoa học dữ liệu là bước tiếp theo của khoa học máy tính.

c) Phân tích dữ liệu là áp dụng mô hình cho dữ liệu đã chuẩn bị để chọn lọc một số yếu tố quan trọng nhằm giải quyết vấn đề.

d) Phân tích dữ liệu là mục đích của khoa học dữ liệu.

Tóm tắt bài học

- Khoa học dữ liệu là lĩnh vực liên ngành, nghiên cứu sử dụng dữ liệu để có những hiểu biết sâu sắc làm cơ sở cho những quyết định mang lại hiệu quả cao.
- Khoa học dữ liệu đã có đóng góp quan trọng trong một số thành tựu khoa học như: dự án Bộ gen người HGP, các dự án nghiên cứu và khám phá không gian vũ trụ, hệ thống Giám sát đánh bắt cá toàn cầu, các mô hình ngôn ngữ lớn, mô hình phát hiện gian lận của American Express,...