1. Khoa học dữ liệu:

a) Khái niệm khoa học dữ liệu:

Khoa học dữ liệu là một lĩnh vực liên ngành kết hợp các phương pháp từ toán học, thống kê, và khoa học máy tính với kiến thức chuyên môn trong các lĩnh vực ứng dụng như kinh doanh, tài chính, y tế, giáo dục. Mục tiêu của Khoa học dữ liệu bao gồm:

1. Phân tích và trực quan hóa dữ liệu: Xem xét các mẫu, xu hướng để hiểu và biểu diễn dữ liệu một cách trực quan nhằm phát hiện vấn đề cần giải quyết.

2. Xây dựng mô hình dự đoán và dự báo: Sử dụng dữ liệu để dự đoán sự kiện tương lai như thay đổi doanh số hoặc biến động về khách hàng.

3. Tối ưu hóa quyết định: Điều chỉnh quyết định dựa trên dữ liệu và sử dụng thuật toán tối ưu hóa để đưa ra quyết định tốt nhất.

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

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

1.Xác định vấn đề: Hiểu rõ vấn đề cần giải quyết để xác định các giả thuyết cần kiểm tra và đánh giá.

2. Thu thập dữ liệu: Thu thập dữ liệu liên quan từ nhiều nguồn, thường là dữ liệu lớn.

3. Chuẩn bị dữ liệu: Lựa chọn, tích hợp, làm sạch dữ liệu, xử lý các giá trị thiếu và chính xác, và chuẩn bị dữ liệu cho phân tích.

4. Phân tích và khai phá dữ liệu: Áp dụng mô hình để chọn lọc yếu tố quan trọng, tìm ra mối quan hệ và quy luật trong dữ liệu để xây dựng mô hình dự báo và phát triển kiến thức.

5.Đánh giá và giải thích: Đánh giá chất lượng mô hình bằng các tiêu chí cụ thể và giải thích tác động của mô hình đối với tổ chức.

6.Ra quyết định và triển khai: Trình bày kết quả phân tích cho cấp lãnh đạo để ra quyết định và triển khai thực tế.

Lý thuyết Tin 12 Cánh diều Bài 2: Giới thiệu về khoa học dữ liệu

2. Một số thành tựu của Khoa học dữ liệu:

- Tài chính – Ngân hàng: - Đánh giá rủi ro.Phát hiện gian lận. Lập mô hình đầu tư.Phân khúc khách hàng.

- Chăm sóc sức khoẻ:Dự đoán dịch bệnh.Cải thiện chất lượng chăm sóc bệnh nhân.Quản lý dịch vụ y tế.Chế tạo thuốc chữa bệnh.

- Sản xuất – Kinh doanh:Ra quyết định tầm chiến lược.Tối ưu hóa quy trình sản xuất.Cá nhân hóa trải nghiệm khách hàng.Đưa ra khuyến nghị cho khách hàng.

- Dịch vụ công nghệ thông tin:Tối ưu hóa hệ thống thông tin.Đảm bảo an ninh mạng.

- Trí tuệ nhân tạo:Phát triển các hệ thống như trợ lý ảo.Thu thập và phân tích dữ liệu lớn để xây dựng và huấn luyện mô hình hiệu quả.

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 (Human Genome Project - HGP):

-Thời gian và Chi phí: Kéo dài 13 năm (1990-2003), tiêu tốn khoảng 3 tỉ USD.

- Mục tiêu: Nghiên cứu cấu trúc và chức năng của các gen trong bộ gen người, xác định các biến thể di truyền và mối quan hệ giữa đột biến và đặc điểm sinh học.

- Kỹ thuật: Lập bản đồ gen và giải trình tự gen.

- Dữ liệu: Bộ gen người có khoảng 3 tỉ cặp base. Giải trình tự một bộ gen sinh ra khoảng 100 gigabyte dữ liệu; giải trình tự nhiều bộ gen có thể tạo ra hàng trăm petabyte dữ liệu.

- Ứng dụng: Phân tích giúp hiểu rõ cách thức hoạt động của gen, chức năng của gen, và mối quan hệ giữa gen với sức khoẻ và bệnh tật.

Trong đại dịch Covid-19:-Các nhà nghiên cứu đã sử dụng máy giải trình tự gen tiên tiến để xác định virus SARS-CoV-2.

- Điều này giúp phân tích cách virus gây bệnh, phát triển phương pháp chẩn đoán, điều trị và phòng ngừa hiệu quả.

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

- Kính thiên văn Kepler: Trong 9 năm hoạt động, thu thập khoảng 678 GB dữ liệu về độ sáng của khoảng 150 nghìn ngôi sao.

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

Ứng dụng của học máy:

- Phát triển thuật toán học máy: Để phân tích lượng dữ liệu khổng lồ, giúp phân loại đặc trưng của hành tinh, phát hiện thay đổi ánh sáng ngôi sao và suy luận về các hành tinh khác trong hệ ngôi sao.

Kết quả:

- Phát hiện: Khoảng 3.200 hệ hành tinh quay quanh các ngôi sao trong dải Ngân Hà, trong đó có khoảng 63 hành tinh có khả năng duy trì sự sống.

c) Hệ thống Giám sát đánh bắt cá toàn cầu:

Hệ thống Giám sát đánh bắt cá toàn cầu (Global Fishing Watch):

- Mục tiêu: Ngăn chặn đánh bắt cá bất hợp pháp.

- Công nghệ sử dụng: Dữ liệu vệ tinh từ hệ thống của Google.

- Hoạt động: Ghi lại hàng triệu vị trí của tàu cá trên toàn thế giới mỗi ngày. Xác định mục đích chuyến đi và điểm xuất phát của tàu.

- Kết quả: Theo dõi hoạt động đánh bắt cá theo thời gian thực, giúp phát hiện và xác định tàu đánh bắt cá bất hợp pháp.

d) Các mô hình ngôn ngữ lớn:

Các mô hình ngôn ngữ lớn (LLM):

- Khái niệm: Mô hình AI đặc biệt để hiểu ngôn ngữ tự nhiên.

- Ví dụ nổi tiếng: GPT-3 với 175 tỉ tham số.

- Đặc điểm: Số lượng tham số lớn giúp mô hình hiểu và xử lý ngôn ngữ tinh vi hơn.

- Thành tựu: GPT-3 đạt được kết quả ấn tượng trong xử lý ngôn ngữ tự nhiên, có thể thực hiện các nhiệm vụ phức tạp với hiệu suất cao, đôi khi vượt qua khả năng của con người.

e) Mô hình phát hiện gian lận của American Express:

- Thành công: American Express đã cải thiện khả năng phát hiện gian lận bằng cách áp dụng khoa học dữ liệu.

- Năm 2014: Triển khai mô hình học máy, nâng cao hiệu quả phát hiện gian lận lên 30% so với hệ thống cũ.

- Năm 2017: Phát triển công cụ xác thực sinh trắc học, giảm 60% giao dịch gian lận.

- Hiện tại: Đạt tỉ lệ gian lận thấp nhất ở Mỹ trong 14 năm liên tiếp theo báo cáo của Nilson tháng 2 năm 2021