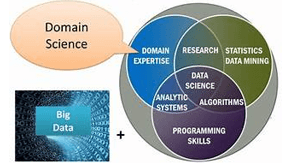
# Bài 26: Làm quen với Khoa học dữ liệu

**Giải Tin học 12 Bài 26: Làm quen với Khoa học dữ liệu**  
**Khởi động trang 140 Tin học 12**: Những năm gần đây, cùng với AI, Khoa học dữ liệu (data science) đã trở thành lĩnh vực thu hút sự quan tâm đặc biệt trên toàn thế giới. Hãy nhập từ khoá "data science” vào thanh công cụ tìm kiếm Google và cho nhận xét về kết quả tìm kiếm mà em nhận được.  
**Lời giải:**  
Thông tin trên internet:  
  
Khoa học dữ liệu (Data Science) là một lĩnh vực đa ngành sử dụng thống kê, tính toán khoa học, phương pháp khoa học, thuật toán và hệ thống để trích xuất hoặc suy luận kiến thức và thông tin từ dữ liệu có thể không chính xác, có cấu trúc hoặc không có cấu trúc 1 2 Đây là một lĩnh vực đang phát triển mạnh mẽ và có ảnh hưởng lớn đến nhiều ngành công nghiệp.  
Cụ thể, khoa học dữ liệu kết hợp các yếu tố sau:  
- Toán học và thống kê: Sử dụng các phương pháp thống kê và toán học để phân tích dữ liệu và tạo ra các mô hình dự đoán.  
- Lập trình chuyên sâu: Sử dụng các ngôn ngữ lập trình như R hoặc Python để xử lý và phân tích  
dữ liệu.  
- Trí tuệ nhân tạo (AI) và học máy: Áp dụng các thuật toán học máy để xây dựng mô hình dự đoán và phân loại. Kiến thức về lĩnh vực cụ thể: Hiểu về lĩnh vực mà dữ liệu đang được áp dụng, ví dụ: y học, tài chính, marketing, v.v.  
Với sự gia tăng về khối lượng dữ liệu và nguồn dữ liệu đa dạng, khoa học dữ liệu đã trở thành một trong những lĩnh vực phát triển nhanh nhất ở mọi ngành công nghiệp. Không ngạc nhiên khi vai trò của nhà khoa học dữ liệu đã được gọi là “công việc hấp dẫn nhất của thế kỷ 21" bởi Harvard Business Review 1. Các tổ chức ngày càng phụ thuộc vào họ để hiểu và phân tích dữ liệu, từ đó đưa ra các đề xuất hành động để cải thiện kết quả kinh doanh.  
Quá trình khoa học dữ liệu bao gồm các giai đoạn sau:  
1. Thu thập dữ liệu: Thu thập dữ liệu từ nhiều nguồn khác nhau, bao gồm cả dữ liệu có cấu trúc và không có cấu trúc.  
2. Lưu trữ và xử lý dữ liệu: Xử lý dữ liệu, chuẩn hóa, biến đổi và lưu trữ vào các hệ thống dữ liệu như data warehouse hoặc data lake.  
3. Phân tích dữ liệu: Khám phá dữ liệu, tìm hiểu các mẫu và xu hướng, và xây dựng các mô hình dự đoán.  
4. Trình bày kết quả: Trình bày thông tin dưới dạng báo cáo và biểu đồ để giúp người quản lý và ra quyết định hiểu rõ hơn về dữ liệu và tác động của nó lên kết quả kinh doanh 1.  
Tóm lại, khoa học dữ liệu không chỉ là một công cụ hữu ích, mà còn là một phần quan trọng của sự phát triển và ứng dụng thực tế.  
Với sự gia tăng về khối lượng dữ liệu và nguồn dữ liệu đa dạng, khoa học dữ liệu đã trở thành một trong những lĩnh vực phát triển nhanh nhất ở mọi ngành công nghiệp. Không ngạc nhiên khi vai trò của nhà khoa học dữ liệu đã được gọi là "công việc hấp dẫn nhất của thế kỷ 21" bởi Harvard Business Review 1. Các tổ chức ngày càng phụ thuộc vào họ để hiểu và phân tích dữ liệu, từ đó đưa ra các đề xuất hành động để cải thiện kết quả kinh doanh.  
Quá trình khoa học dữ liệu bao gồm các giai đoạn sau:  
1. Thu thập dữ liệu: Thu thập dữ liệu từ nhiều nguồn khác nhau, bao gồm cả dữ liệu có cấu trúc và không có cấu trúc.  
2. Lưu trữ và xử lý dữ liệu: Xử lý dữ liệu, chuẩn hóa, biến đối và lưu trữ vào các hệ thống dữ liệu như data warehouse hoặc data lake.  
3. Phân tích dữ liệu: Khám phá dữ liệu, tìm hiểu các mẫu và xu hướng, và xây dựng các mô hình dự đoán.  
4. Trình bày kết quả: Trình bày thông tin dưới dạng báo cáo và biểu đồ để giúp người quản lý và ra quyết định hiểu rõ hơn về dữ liệu và tác động của nó lên kết quả kinh doanh 1.  
Tóm lại, khoa học dữ liệu không chỉ là một công cụ hữu ích, mà còn là một phần quan trọng của sự phát triển và ứng dụng thực tế.  
  
**Hoạt động 1 trang 140 Tin học 12**: Có thể hiểu đơn giản Khoa học dữ liệu là lĩnh vực khoa học nghiên cứu về dữ liệu. Như vậy, đối tượng nghiên cứu của Khoa học dữ liệu chính là dữ liệu. Theo em, Khoa học dữ liệu không bao gồm công việc nào sau đây?  
A. Nghiên cứu phát triển các phương pháp thu thập và quản lí dữ liệu.  
B. Khai phá các thông tin, tri thức từ dữ liệu thu được để nâng cao hiệu quả kinh doanh, quản lí.  
C. Kinh doanh, phân phối dữ liệu thu thập được cho các cá nhân, tổ chức quan tâm.  
D. Phát triển và áp dụng các phương pháp và kĩ thuật để nhận biết các mẫu hình, các quan hệ và xu hướng có trong dữ liệu.  
**Lời giải:**  
Đáp án đúng là: C. Kinh doanh, phân phối dữ liệu thu thập được cho các cá nhân, tổ chức quan tâm.  
Lý do chọn đáp án này là vì các hoạt động trong lựa chọn C không nằm trong phạm vi hoạt động của Khoa học Dữ liệu. Dưới đây là lý do chi tiết:  
A. Nghiên cứu phát triển các phương pháp thu thập và quản lí dữ liệu: Đây là một phần quan trọng của Khoa học Dữ liệu. Việc thu thập và quản lý dữ liệu một cách hiệu quả là cần thiết để có thể sử dụng dữ liệu đó cho các mục đích nghiên cứu và ứng dụng.  
B. Khai phá các thông tin, tri thức từ dữ liệu thu được để nâng cao hiệu quả kinh doanh, quản lí: Đây cũng là một phần chính của Khoa học Dữ liệu. Việc khai thác tri thức từ dữ liệu có thể giúp cải thiện quyết định kinh doanh và quản lý thông qua việc phân tích dữ liệu và rút ra các insights quan trọng.  
C. Kinh doanh, phân phối dữ liệu thu thập được cho các cá nhân, tổ chức quan tâm: Đây không phải là một phần của Khoa học Dữ liệu. Trong thực tế, việc kinh doanh và phân phối dữ liệu thường liên quan đến các hoạt động thương mại, không phải nghiên cứu và phân tích dữ liệu.  
D. Phát triển và áp dụng các phương pháp và kỹ thuật để nhận biết các mẫu hình, các quan hệ và xu hướng có trong dữ liệu: Đây là một phần quan trọng của Khoa học Dữ liệu. Việc phát triển và áp dụng các phương pháp và kỹ thuật như machine learning và data mining giúp phân tích và nhận biết các mẫu hình và xu hướng trong dữ liệu.  
**Câu hỏi 1 trang 142 Tin học 12**: Học máy và tri thức chuyên ngành có vai trò gì trong Khoa học dữ liệu?  
**Lời giải:**  
Học máy và tri thức chuyên ngành đóng vai trò quan trọng trong Khoa học dữ liệu:  
- Học máy:  
+ Cung cấp các công cụ và kỹ thuật để xử lý, phân tích và khai phá dữ liệu.  
+ Sử dụng ngôn ngữ lập trình và thuật toán học máy để xây dựng các mô hình dự đoán.  
+ Đóng vai trò quan trọng trong việc phân tích và khai phá dữ liệu.  
- Tri thức chuyên ngành:  
+ Là tri thức của từng lĩnh vực, ví dụ kinh doanh, y tế, khoa học xã hội.  
+ Giúp hiểu ngữ cảnh và ý nghĩa của dữ liệu.  
+ Đánh giá chất lượng và độ chính xác của dữ liệu, diễn giải kết quả phân tích và khai phá dữ liệu theo cách phù hợp với lĩnh vực ứng dụng để đưa ra quyết định đúng đắn.  
  
**Câu hỏi 2 trang 142 Tin học 12**: Tính chất liên ngành của Khoa học dữ liệu được thể hiện như thế nào?  
**Lời giải:**  
Tính chất liên ngành của Khoa học dữ liệu được thể hiện qua việc kết hợp các yếu tố từ nhiều lĩnh vực khác nhau:  
- Kết hợp của nhiều lĩnh vực: Khoa học dữ liệu sử dụng kiến thức từ khoa học máy tính, toán học, thống kê, và tri thức chuyên ngành. Điều này giúp tạo ra một hệ thống toàn diện và đa dạng để xử lý và phân tích dữ liệu.  
- Sự đa dạng trong ứng dụng: Khoa học dữ liệu áp dụng trong nhiều lĩnh vực khác nhau như kinh doanh, y tế, khoa học xã hội, marketing, và nhiều ngành công nghiệp khác. Sự đa dạng này thể hiện tính liên ngành của nó.  
- Hiểu biết về lĩnh vực cụ thể: Tri thức chuyên ngành là yếu tố quan trọng để hiểu ngữ cảnh và ý nghĩa của dữ liệu. Khoa học dữ liệu không chỉ là việc xử lý dữ liệu mà còn đòi hỏi hiểu biết sâu về lĩnh vực mà dữ liệu đang áp dụng.  
  
**Hoạt động 2 trang 142 Tin học 12**: Khi nói tới dữ liệu lớn người ta thường nghĩ tới kích thước lớn của dữ liệu. Tuy nhiên, trong thực tế, có những dữ liệu không chỉ có kích thước lớn, thường xuyên được cập nhật mà còn bao gồm nhiều loại khác nhau. Em có thể chỉ ra một vài ví dụ về những dữ liệu như vậy không?  
**Lời giải:**  
Ví dụ:  
- Dữ liệu tài chính: Dữ liệu từ thị trường chứng khoán, giao dịch ngân hàng, dữ liệu tài chính của các công ty, và dữ liệu về tiền tệ.  
- Dữ liệu xã hội (Social Media): Dữ liệu từ các trang web xã hội như Facebook, Twitter, Instagram, LinkedIn, và YouTube. Đây bao gồm các bài viết, bình luận, hình ảnh, video, và thông tin cá nhân của người dùng.  
- Dữ liệu y tế: Dữ liệu từ bệnh viện, phòng khám, dữ liệu về bệnh nhân, dược phẩm, và nghiên cứu y học.  
- Dữ liệu địa lý và vận tải: Dữ liệu từ hệ thống định vị toàn cầu (GPS), thông tin về giao thông, thời tiết, và hành trình vận chuyển.  
- Dữ liệu sản xuất và công nghiệp: Dữ liệu từ máy móc, cảm biến, quá trình sản xuất, và quản lý chuỗi cung ứng.  
**Câu hỏi 1 trang 144 Tin học 12**: Giới thiệu một vài thành tựu của Khoa học dữ liệu mà em tâm đắc nhất.  
**Lời giải:**  
Dưới đây là một vài thành tựu của Khoa học dữ liệu mà em tâm đắc nhất:  
- Đổi mới quá trình ra quyết định: khoa học dữ liệu giúp cải thiện quá trình ra quyết định bằng cách phân tích dữ liệu và đưa ra thông tin hữu ích. Điều này ảnh hưởng đến nhiều khía cạnh của cuộc sống và kinh doanh.  
- Tự động hoá: khoa học dữ liệu giúp tự động hoá nhiều tác vụ, từ việc xử lý dữ liệu đến việc tối ưu hóa quyết định. Điều này giúp tiết kiệm thời gian và tối ưu hiệu suất.  
- Cá nhân hoá dịch vụ và cải thiện trải nghiệm khách hàng: dựa trên dữ liệu, chúng ta có thể cá nhân hoá dịch vụ và sản phẩm để đáp ứng nhu cầu của từng khách hàng một cách tốt nhất. Điều này cải thiện trải nghiệm của khách hàng và tạo sự hài lòng.  
  
**Câu hỏi 2 trang 144 Tin học 12**: Quan sát Hình 26.4 và cho biết kết quả khuyến nghị là gì.  
  
**Lời giải:**  
Kết quả khuyến nghị là các bài hát của Estas Tonne.  
  
**Luyện tập 1 trang 144 Tin học 12**: Tại sao lại có thể nói Khoa học dữ liệu góp phần tạo ra nhiều cơ hội mới cho các lĩnh vực khác nhau?  
**Lời giải:**  
Khoa học dữ liệu góp phần tạo ra nhiều cơ hội mới cho các lĩnh vực khác nhau vì các lí do sau:  
- Quyết định dựa trên dữ liệu: Khoa học dữ liệu cho phép tổ chức và cá nhân sử dụng dữ liệu để đưa ra quyết định sáng suốt và chính xác hơn. Phân tích và khai phá dữ liệu giúp dự báo và phân tích xu hướng phát triển, từ đó giúp chuẩn bị sẵn sàng và thích nghi với thay đổi, đồng thời đưa ra quyết định kinh doanh phù hợp.  
- Tự động hoá và đổi mới sáng tạo: Khoa học dữ liệu sử dụng các mô hình học máy để tự động hoá các tác vụ lặp đi lặp lại và tốn thời gian. Điều này giúp tiết kiệm thời gian và chi phí, tăng tính hiệu quả và độ chính xác. Các tổ chức có thể áp dụng Khoa học dữ liệu để tự động hoá nhiều quy trình và công việc trong nhiều lĩnh vực, từ sản xuất, dịch vụ khách hàng đến quản lí tài chính.  
- Cá nhân hoá dịch vụ và cải thiện trải nghiệm khách hàng: Khoa học dữ liệu cho phép cá nhân hoá các dịch vụ dựa trên phân tích dữ liệu về khách hàng. Điều này giúp cung cấp thông tin đầy đủ hơn về nhu cầu, sở thích và hành vi của khách hàng, từ đó đưa ra giải pháp cải thiện trải nghiệm khách hàng và tăng doanh số. Các hệ khuyến nghị và các ứng dụng y học cá nhân hoá là ví dụ điển hình.  
  
**Luyện tập 2 trang 144 Tin học 12**: Các tổ chức có thể sử dụng Khoa học dữ liệu để dự đoán thời điểm những trục trặc của thiết bị có thể xảy ra. Hãy phân tích để thấy được, trong trường hợp cụ thể này, Khoa học dữ liệu có thể giúp đổi mới hoàn toàn quy trình bảo trì thiết bị, thay thế quy trình hoạt động chưa hiệu quả.  
**Lời giải:**  
Khoa học dữ liệu có thể giúp đổi mới hoàn toàn quy trình bảo trì thiết bị và thay thế quy trình hoạt động chưa hiệu quả bằng cách dự đoán thời điểm xảy ra trục trặc của thiết bị. Dưới đây là phân tích chi tiết về cách Khoa học dữ liệu có thể đóng góp trong việc đổi mới quy trình bảo trì thiết bị:  
- Dự đoán trục trặc: Khoa học dữ liệu có thể sử dụng các mô hình dự đoán để phân tích dữ liệu từ các thiết bị và đưa ra dự đoán về thời điểm xảy ra trục trặc. Bằng cách phân tích các thông số và mẫu dữ liệu liên quan đến hoạt động của thiết bị, Khoa học dữ liệu có thể nhận biết các dấu hiệu tiền đề của sự cố và đưa ra dự báo về thời gian tiếp theo khi trục trặc có thể xảy ra.  
- Tối ưu hoá lịch trình bảo trì: Dự đoán thời điểm trục trặc giúp tổ chức tối ưu hoá lịch trình bảo trì. Thay vì thực hiện bảo trì theo lịch trình cố định, tổ chức có thể dựa trên dự đoán để lên kế hoạch bảo trì đúng thời điểm khi thiết bị có khả năng gặp sự cố. Điều này giúp tránh tình trạng bảo trì quá sớm hoặc quá trễ, tối ưu hóa sử dụng tài nguyên và giảm thiểu thời gian thiết bị không hoạt động.  
- Phân tích nguyên nhân và tối ưu hiệu suất: Khoa học dữ liệu có thể phân tích dữ liệu từ các thiết bị để xác định nguyên nhân gây ra trục trặc. Thông qua việc phân tích dữ liệu lịch sử và tương quan giữa các yếu tố, Khoa học dữ liệu có thể tìm ra các yếu tố ảnh hưởng và đề xuất các cải tiến để tối ưu hiệu suất và giảm thiểu trục trặc trong tương lai.  
- Tự động hóa quy trình bảo trì: Khoa học dữ liệu có thể hỗ trợ tự động hóa quy trình bảo trì thiết bị. Dựa trên dữ liệu từ các thiết bị và các mô hình học máy, tổ chức có thể phát triển các hệ thống tự động để giám sát, phát hiện và xử lý sự cố một cách tự động. Điều này giúp giảm thiểu sự can thiệp của con người và tăng tính hiệu quả của quy trình bảo trì.  
  
**Vận dụng trang 144 Tin học 12**: Trong thực tế, vẫn có trường hợp dùng các thuật ngữ Khoa học dữ liệu và Phân tích dữ liệu thay thế cho nhau. Điều này không hoàn toàn chính xác. Hãy truy cập Internet để tìm hiểu sự khác biệt giữa hai khái niệm này.  
**Lời giải:**  
Tóm tắt thông tin trên internet:  
Trong thực tế, thuật ngữ "Khoa học dữ liệu" và "Phân tích dữ liệu" không hoàn toàn tương đương và có những khác biệt nhất định. Dưới đây là một trình bày ngắn về sự khác biệt giữa hai khái niệm này:  
- Khoa học dữ liệu (Data Science): Khoa học dữ liệu là lĩnh vực toàn diện và đa ngành, liên quan đến việc thu thập, xử lý, phân tích và hiểu dữ liệu từ nhiều nguồn khác nhau. Khoa học dữ liệu kết hợp các phương pháp từ các lĩnh vực như thống kê, học máy, khai phá dữ liệu và trí tuệ nhân tạo để khám phá thông tin có giá trị từ dữ liệu và đưa ra dự đoán, quyết định hoặc hiểu biết sâu hơn.  
- Phân tích dữ liệu (Data Analytics): Phân tích dữ liệu tập trung vào việc áp dụng các phương pháp thống kê và mô hình hóa để khám phá thông tin từ dữ liệu. Phân tích dữ liệu có thể tập trung vào việc tìm hiểu các mẫu, xu hướng và mối quan hệ trong dữ liệu để đưa ra hiểu biết và giải thích sự kiện đã xảy ra. Nó thường liên quan đến việc sử dụng các phương pháp và công cụ phân tích để trả lời các câu hỏi cụ thể và đưa ra thông tin hữu ích cho quyết định kinh doanh.