# TỔNG QUAN VỀ DỮ LIỆU LỚN

## Khái niệm dữ liệu lớn

Dữ liệu lớn [1] là thuật ngữ dùng để chỉ các tập dữ liệu khổng lồ có cấu trúc lớn, đa dạng và phức tạp, khó lưu trữ, phân tích và trực quan hóa cho các quy trình hoặc kết quả tiếp theo. Quá trình nghiên cứu về lượng dữ liệu khổng lồ để khám phá các mẫu ẩn và mối tương quan bí mật được gọi là phân tích dữ liệu lớn. Những thông tin hữu ích này cho các công ty hoặc tổ chức với sự trợ giúp của việc thu thập thông tin chi tiết phong phú và sâu sắc hơn và có được lợi thế cạnh tranh. Vì lý do này, việc triển khai dữ liệu lớn cần được phân tích và thực thi một cách chính xác nhất có thể.

Dữ liệu lớn có thể là dữ liệu có cấu trúc, bán cấu trúc hoặc không có cấu trúc. Trong đó, khoảng hơn 80% của tất cả các dữ liệu lớn là không có cấu trúc.

Đặc biệt, dữ liệu lớn chủ yếu được thu thập thông qua các công cụ tìm kiếm, nền tảng mạng xã hội, điện thoại di động, tivi thông minh,…. cùng các nguồn khác mà doanh nghiệp có thể truy cập để khai thác.

Đặc trưng 7V của dữ liệu lớn :

* Volume – Khối lượng: Khối lượng của Big data rất lớn, không thể đong đếm bằng đơn vị cụ thể. Với sự phát triển mạnh mẽ của internet, thiết bị di động, mạng xã hội và công nghệ IoT (Internet of Things), khối lượng dữ liệu đang tạo ra tăng đáng kể hàng ngày, hàng giờ.
* Velocity – Tốc độ: Big data có tốc độ xử lý dữ liệu nhanh chóng. Các ứng dụng phổ biến trong lĩnh vực Internet, tài chính, ngân hàng, y tế - chăm sóc sức khỏe,… phần lớn dữ liệu lớn được xử lý real-time (thời gian thực). Công nghệ quản lý dữ liệu lớn ngày một tiên tiến cho phép chúng ta xử lý ngay lập tức trước khi lưu trữ vào cơ sở dữ liệu.
* Variety – Đa dạng: Trước đây, hầu hết dữ liệu có thể được thu thập gọn gàng trong bảng, đó là dữ liệu có cấu trúc. Hiện nay, dữ liệu chủ yếu xuất hiện dưới dạng phi cấu trúc như bài viết trên mạng xã hội, âm thanh, video,… Công nghệ Big Data cho phép liên kết, phân tích đa dạng chủng loại dữ liệu với nhau.
* Value – Giá trị: Giá trị thông tin là tính chất quan trọng nhất của công nghệ Big Data. Ví dụ: Một doanh nghiệp muốn ứng dụng Big Data cho bài toán mô hình hoạt động kinh doanh của mình. Trước tiên, doanh nghiệp cần xác định rõ những giá trị thông tin hữu ích Big Data có thể mang lại trong việc giải quyết nhu cầu của doanh nghiệp.
* Veracity – Độ chính xác: Với sự đa dạng cùng số lượng lớn, độ chính xác là một trong những tính chất phức tạp nhất của Big Data. Tính chất này đề cập đến việc khai phá chất lượng tập dữ liệu và làm sạch dữ liệu đó một cách hệ thống. Từ đó làm tăng độ tin cậy, chính xác, giúp ích cho việc phân tích. Đây cũng là một bước quan trọng của Big Data.
* Variability – Tính khả biến: Quản lý và bối cảnh hóa dữ liệu theo cách cung cấp cấu trúc, ngay cả trong môi trường dữ liệu dễ biến đổi và không thể đoán trước.
* Visualization – Hình dung: Hình dung giúp người dùng có thể hiểu rõ thông tin được phân tích, trích xuất từ kho dữ liệu. Sử dụng biểu đồ, đồ thị để trực quan hóa một lượng lớn dữ liệu phức tạp sẽ hiệu quả hơn nhiều trong việc truyền đạt ý nghĩa so với bảng tính, báo cáo chứa đầy các con số và công thức truyền thống.

## Một số phương pháp phân tích dữ liệu lớn

* **Phân tích cụm (Cluster analysis)**

Phân tích cụm (Cluster analysis) được thực hiện dựa trên việc nhóm các phần dữ liệu có đặc điểm chung với nhau. Vì không có biến đích khi phân nhóm, phương pháp này thường được sử dụng để tìm các mẫu ẩn trong dữ liệu hoặc cung cấp ngữ cảnh bổ sung cho một tập dữ liệu.

Trong kinh doanh, marketing, các doanh nghiệp vẫn thường áp dụng phương pháp phân tích cụm nhằm nhóm các tập khách hàng dựa trên nhân khẩu học, hành vi mua hàng, khả năng tài chính,… Mục tiêu là mang đến những trải nghiệm tốt nhất dựa trên nhu cầu, thị hiếu của một lượng lớn khách hàng, trong khi vẫn đảm bảo tối ưu hóa nguồn lực vận hành.

* **Phân tích tổ hợp (Cohort analysis)**

Phương pháp này sử dụng dữ liệu lịch sử để kiểm tra và đối chiếu một phân khúc xác định về hành vi của người dùng, sau đó nhóm chúng lại với những phân khúc khác có đặc điểm tương tự. Bằng phương pháp này, bạn có thể hiểu được nhu cầu của người tiêu dùng, thậm chí là một nhóm đối tượng mục tiêu số lượng lớn hơn.

Về ứng dụng, phân tích tổ hợp giúp các Marketers hiểu được tác động của chiến dịch trên các nhóm khách hàng cụ thể. Ví dụ, hãy tưởng tượng bạn gửi một chiến dịch email khuyến khích khách hàng đăng ký trang web. Bạn tạo hai phiên bản của chiến dịch với các thiết kế, CTA (Call to action) và nội dung quảng cáo khác nhau. Sau đó, bạn có thể sử dụng phân tích tổ hợp để theo dõi hiệu suất của chiến dịch trong một khoảng thời gian dài hơn và hiểu loại nội dung nào đang thúc đẩy khách hàng đăng ký, mua hàng hoặc tương tác theo những cách khác.

* **Phân tích hồi quy (Regression analysis)**

Hồi quy sử dụng dữ liệu lịch sử để hiểu tác động đến giá trị của biến phụ thuộc khi một (hồi quy tuyến tính) hoặc nhiều biến độc lập (hồi quy bội) thay đổi hoặc giữ nguyên. Bằng cách hiểu mối quan hệ của từng biến và cách chúng phát triển trong quá khứ, bạn có thể dự đoán các kết quả có thể xảy ra và đưa ra quyết định tốt hơn trong tương lai.

* **Mạng nơron**

Mạng nơ-ron là nền tảng cho các thuật toán thông minh của học máy. Nó là một dạng phân tích có sự can thiệp tối thiểu, để hiểu cách bộ não con người tạo ra những hiểu biết sâu sắc và dự đoán các giá trị. Mạng nơ-ron học hỏi từ mọi dữ liệu, nghĩa là chúng phát triển và tiến bộ theo thời gian.

* **Phân tích nhân tố (Factor analysis)**

Phân tích nhân tố còn được gọi là “giảm chiều dữ liệu” (dimension reduction) mô tả sự biến thiên của những biến có tương quan được quan sát bằng một số nhỏ hơn các biến không quan sát được gọi là nhân tố. Mục đích ở đây là phát hiện ra các biến tiềm ẩn độc lập.

Một ví dụ về phương pháp phân tích dữ liệu này là đánh giá của khách hàng về sản phẩm. Đánh giá ban đầu dựa trên các biến số khác nhau như màu sắc, hình dạng, chất liệu, sự thoải mái, cửa hàng, tần suất sử dụng. Trong trường hợp này, danh sách các biến số có thể rất dài, tùy thuộc vào những gì bạn muốn theo dõi. Do đó, phân tích nhân tố đưa ra bức tranh tổng quát bằng cách tóm tắt tất cả các biến này thành các nhóm đồng nhất, ví dụ, bằng cách nhóm các biến màu sắc, vật liệu, chất lượng và xu hướng thành một biến tiềm ẩn của thiết kế.

* **Khai thác dữ liệu (Data mining)**

Khai thác dữ liệu là phương pháp phân tích dữ liệu giúp trích xuất thông tin từ một tập dữ liệu nhằm xác định xu hướng, mẫu và dữ liệu hữu ích. Cùng với phân tích dự đoán, khai thác dữ liệu là một nhánh của khoa học thống kê sử dụng các thuật toán phức tạp, không chỉ bao hàm bước phân tích thô, mà còn liên quan tới cơ sở dữ liệu, quản lý dữ liệu, tiền xử lý dữ liệu, suy luận thống kê,…

* **Phân tích văn bản (Text analysis)**

Phân tích văn bản, còn được gọi là khai thác văn bản, hoạt động bằng cách lấy các bộ dữ liệu văn bản lớn và sắp xếp chúng nhằm dễ quản lý hơn. Phương pháp này giúp bạn trích xuất dữ liệu thực sự liên quan đến tổ chức của mình và sử dụng dữ liệu đó nhằm phát triển những thông tin hữu ích phục vụ việc ra quyết định. Ví dụ: việc phân tích dữ liệu từ nhiều  nguồn văn bản khác nhau như bài viết đánh giá sản phẩm trên mạng xã hội hoặc phản hồi khảo sát giúp hiểu sâu sắc hơn về đối tượng mục tiêu, từ đó cho phép xây dựng các chiến dịch, dịch vụ đáp ứng nhu cầu của khách hàng tiềm năng.

Nhờ sự kết hợp của học máy và các thuật toán thông minh, phân tích văn bản hiện nay còn cho phép thực hiện các quy trình phân tích nâng cao như phân tích cảm xúc. Phân tích cảm xúc thường được sử dụng để theo dõi danh tiếng của thương hiệu và sản phẩm cũng như để hiểu mức độ thành công của trải nghiệm khách hàng.

* **Phân tích chuỗi thời gian (Time series analysis)**

Phân tích chuỗi thời gian được sử dụng để phân tích một tập hợp dữ liệu thu thập trong một khoảng thời gian xác định. Ngoài ra, nó còn cho phép các nhà nghiên cứu biết được liệu các biến có thay đổi trong suốt thời gian nghiên cứu hay không, các biến khác nhau phụ thuộc như thế nào và nó đạt được kết quả cuối cùng ra sao.

Trong kinh doanh, phương pháp này được sử dụng để hiểu nguyên nhân của các xu hướng và mô hình khác nhau, từ đó rút ra những hiểu biết có giá trị. Đồng thời, phương pháp này cũng có thể kết hợp với dự báo chuỗi thời gian nhằm dự báo sự kiện có thể xảy ra trong tương lai.

* **Cây quyết định (Decision Trees)**

Phân tích dựa trên cây quyết định hoạt động như một công cụ hỗ trợ để đưa ra các quyết định chiến lược và thông minh. Bằng cách hiển thị trực quan các kết quả, hậu quả và chi phí tiềm năng trong mô hình dạng cây, các nhà nghiên cứu và người dùng doanh nghiệp có thể dễ dàng đánh giá tất cả các yếu tố liên quan và chọn cách hành động tốt nhất. Cây quyết định thường được dùng để phân tích dữ liệu định lượng, nó cho phép cải thiện quy trình ra quyết định bằng cách giúp bạn xác định các cơ hội cải tiến, giảm chi phí, nâng cao hiệu quả hoạt động và sản xuất.

Cụ thể, phương pháp này hoạt động giống như một sơ đồ bắt đầu với quyết định chính mà bạn cần đưa ra và phân nhánh dựa trên các kết quả và hậu quả khác nhau của mỗi quyết định. Mỗi kết quả sẽ nêu ra những hậu quả, chi phí và lợi ích của riêng nó và khi kết thúc phân tích, bạn có thể so sánh từng kết quả và đưa ra quyết định thông minh nhất.

* **Phân tích thuộc tính (Conjoint analysis)**

Phân tích thuộc tính thường được sử dụng trong các cuộc khảo sát để hiểu cách người dùng đánh giá các thuộc tính khác nhau của một sản phẩm hoặc dịch vụ. Ví dụ, khi nói đến việc mua hàng, một số khách hàng có thể tập trung vào giá, những người khác tập trung vào tính năng, hay tính bền vững của sản phẩm. Bạn có thể tìm thấy các thuộc tính này bằng phân tích kết hợp. Như vậy, các công ty có thể xác định chiến lược giá cả, tùy chọn gói sản phẩm, dịch vụ,….

## Ứng dụng của phân tích dữ liệu lớn

Một số lĩnh vực đã và đang đưa phân tích dữ liệu lớn vào ứng dụng có thể kể đến:

* Ngân hàng: Big Data đã và đang được ứng dụng hiệu quả với các hoạt động: từ thu tiền mặt đến quản lý tài chính. Từ đó, giúp giảm bớt rắc rối của khách hàng và gia tăng doanh thu cho các ngân hàng.

Cách thức ứng dụng Big Data trong ngân hàng:

* Sử dụng các kỹ thuật phân cụm giúp ngân hàng đưa ra quyết định quan trọng. Hệ thống phân tích có thể xác định các địa điểm chi nhánh nơi tập trung nhiều nhu cầu của khách hàng, để đề xuất lập chi nhánh mới.
* Quy tắc kết hợp để dự đoán lượng tiền mặt cần thiết sẵn sàng cung ứng ở một chi nhánh tại thời điểm cụ thể.
* Khoa học dữ liệu là nền tảng của hệ thống ngân hàng kĩ thuật số.
* Học máy và AI được nhiều ngân hàng sử dụng để phát hiện các hoạt động gian lận và báo cáo cho các chuyên viên liên quan.
* Khoa học dữ liệu hỗ trợ xử lý, lưu trữ và phân tích lượng dữ liệu khổng lồ từ các hoạt động hàng ngày và giúp đảm bảo an ninh cho ngân hàng.
* Y tế: Với Big Data, người dùng có thể xác định phương hướng điều trị, cải thiện quá trình chăm sóc sức khỏe, giảm lãng phí tiền bạc và thời gian.

Cách thức ứng dụng của Big Data trong ngành y tế:

* Cho phép người quản lý ca dự đoán các bác sĩ cần thiết vào những thời điểm cụ thể
* Theo dõi tình trạng bệnh nhân bằng hồ sơ sức khỏe điện tử.
* Sử dụng các thiết bị kỹ thuật số có thể đeo, từ đó, hệ thống Big Data có thể theo dõi tình trạng bệnh nhân, gửi báo cáo cho các bác sĩ.
* Big Data có thể đánh giá các triệu chứng và xác định nhiều bệnh ở giai đoạn đầu.
* Có thể lưu giữ các hồ sơ nhạy cảm được bảo mật và lưu trữ lượng dữ liệu khổng lồ một cách hiệu quả.
* Ứng dụng Big Data có thể báo trước khu vực có nguy cơ bùng phát dịch như: sốt xuất huyết hoặc sốt rét.
* Thương mại điện tử: Big Data có thể tạo lợi thế cạnh tranh cho doanh nghiệp bằng cách cung cấp thông tin chuyên sâu và các bản báo cáo phân tích xu hướng tiêu dùng.

Cách thức ứng dụng của Big Data trong thương mại điện tử:

* Có thể thu thập dữ liệu và yêu cầu của khách hàng
* Tạo ra một mô hình tiếp thị hiệu suất cao.
* Nhà quản lý có thể xác định các sản phẩm được xem nhiều nhất và tối ưu thời gian hiển thị của các trang sản phẩm này.
* Đánh giá hành vi của khách hàng và đề xuất các sản phẩm tương tự. Điều này làm tăng khả năng bán hàng, từ đó tạo ra doanh thu cao hơn.
* Nếu bất kỳ sản phẩm nào được thêm vào giỏ hàng nhưng khách hàng không mua, Big Data có thể tự động gửi code khuyến mại cho khách hàng cụ thể đó.
* Các ứng dụng Big Data còn có thể tạo một báo cáo theo các tiêu chí: độ tuổi, giới tính, địa điểm của khách truy cập, v.v.
* Giáo dục: Big Data tạo ra các phương pháp tiếp cận dựa trên dữ liệu để dạy học sinh.

Cách thức ứng dụng Big Data vào giáo dục:

* Có thể lưu trữ, quản lý, phân tích các bộ dữ liệu lớn bao gồm hồ sơ của sinh viên.
* Duy trì bảo mật bằng cách sử dụng hệ thống quản lý big data có khả năng trích xuất phân cấp.
* Big Data giúp các bộ đề kiểm tra gần như không thể bị lộ
* Big Data cung cấp dữ liệu về các hoạt động trong lớp và giúp đưa ra quyết định cho giáo viên hay người điều hành tổ chức.
* Big Data có thể đánh giá biểu cảm khuôn mặt và di chuyển của học sinh trong lớp từ đó giáo viên có thể đánh giá được chất lượng giảng dạy bằng máy ảnh độ phân giải cao, cảnh quay video và xử lý hình ảnh.
* Cloud (Điện toán đám mây): Với cơ sở dữ liệu quá lớn thì đám mây là phương tiện được sử dụng để cung cấp cơ sở hạ tầng cần thiết cho việc tính toán dữ liệu lớn. Trong cuộc sống thực, nhiều tổ chức đang kết hợp hai công nghệ này để cải thiện hoạt động điều phối kinh doanh của mình.

Cách thức ứng dụng Big Data:

* Cloud cung cấp các công cụ để trích xuất dữ liệu. Từ đó, bằng việc phân tích dữ liệu lớn, thì mục tiêu và các quyết định của kinh doanh được xác định.
* Ứng dụng Big Data trong kinh doanh trên nền tảng Cloud giúp các tổ chức quản lý hiệu quả nhiều công cụ phần mềm và phần cứng.
* Cloud đã tăng tốc độ quản lý và truy cập cơ sở dữ liệu chứa hàng lượng lớn hồ sơ.
* Dịch vụ khách hàng: Để thành công thì doanh nghiệp cần làm hài lòng khách hàng và hiểu nhu cầu của họ. Trong giai đoạn mới gia nhập thị trường và tìm lợi thế cạnh tranh, thì sẽ rất khó để bạn biết khách hàng đang tìm kiếm điều gì. Big Data sẽ giúp chủ doanh nghiệp tìm ra giải pháp và đưa ra đề xuất tốt nhất.

Cách thức ứng dụng của Big Data vào dịch vụ khách hàng:

* Xác định các yêu cầu của khách hàng, tập trung thực hiện nhu cầu của họ.
* Phân tích hành vi, sự quan tâm của khách hàng tạo ra các sản phẩm hướng đến khách hàng.
* Có thể thu thập nhiều dữ liệu về hành vi khách hàng để thiết kế mô hình tiếp thị tối ưu.
* Tìm ra sự tương đồng giữa khách hàng và nhu cầu của họ, từ đó, việc nhắm mục tiêu các chiến dịch quảng cáo có thể chính xác và đạt hiệu quả cao.
* Ngành bán lẻ: Big Data mang lại cơ hội cho lĩnh vực bán lẻ bằng việc giúp xác định hành trình trải nghiệm, xu hướng mua sắm và sự hài lòng của khách hàng từ cách thu thập dữ liệu đa dạng. Từ những dữ liệu thu thập được có thể cải thiện hiệu suất và hiệu quả bán hàng.

Cách thức ứng dụng Big Data trong bán lẻ:

* Big data giúp nhà quản lý xây dựng mô hình chi tiêu của từng khách hàng
* Với các phân tích dự đoán, ngành công nghiệp có thể so sánh tỷ lệ cung – cầu và có thể tránh tung ra sản phẩm không được đón nhận
* Ngành bán lẻ có thể xác định vị trí bố trí sản phẩm trên kệ hàng tùy vào thói quen và nhu cầu mua hàng của khách hàng và thiết lập chiến lược kinh doanh để cải thiện
* Kết hợp phân tích cùng lúc các dữ liệu về thời điểm, giao dịch, truyền thông xã hội, dự báo thời tiết để xác định chính xác sản phẩm phù hợp để cung ứng cho khách hàng.
* Digital Marketing: Digital Marketing là chìa khóa để doanh nghiệp tiếp cận gần nhất với khách hàng. Quan trọng, với Digital Marketing, tất cả các doanh nghiệp ở mọi quy mô lớn nhỏ... đều có thể tiến hành các hoạt động quảng cáo tiếp thị trên các nền tảng truyền thông xã hội..

Cách thức ứng dụng Big Data trong Digital Marketing:

* Phân tích thị trường, đối thủ cạnh tranh và đánh giá mục tiêu kinh doanh. Điều này giúp xác định cơ hội tốt để tiếp tục tiến hành các kế hoạch kinh doanh tiếp theo
* Có thể xác định người dùng trên các phương tiện truyền thông xã hội và nhắm mục tiêu cho họ dựa trên nhân khẩu học, giới tính, thu nhập, tuổi tác và sở thích
* Tạo báo cáo cho chiến dịch quảng cáo:hiệu suất, khách hàng và giải pháp để tạo kết quả tốt hơn
* Khoa học dữ liệu được sử dụng cho các khách hàng nhắm mục tiêu và nuôi dưỡng chu trình khách hàng
* Tập trung vào các chủ đề được tìm kiếm cao và tư vấn cách để nội dung để xếp hạng trang web doanh nghiệp cao hơn trên google (SEO).
* Có thể tạo đối tượng tương tự bằng cách sử dụng cơ sở dữ liệu đối tượng hiện có để nhắm mục tiêu các khách hàng tương tự và kiếm được lợi nhuận.
* Truyền thông & Giải trí: Các công ty truyền thông cần thúc đẩy chuyển đổi kỹ thuật số để phân phối sản phẩm và nội dung của họ nhanh nhất có thể tại thị trường hiện tại. Đối với những người ảnh hưởng trên các phương tiện truyền thông đại chúng, Big Data có thể giúp tìm ra quan điểm hoặc lượt thích của một nghệ sĩ để đo lường mức độ phổ biến trong lĩnh vực truyền thông số.

Cách thức ứng dụng Big Data trong truyền thông và giải trí:

* Giúp thu thập thông tin và nhu cầu của cá nhân
* Xác định thiết bị và thời gian tạo hiệu quả cao nhất thông qua các dữ liệu để phân tích
* Có thể xác định lý do đăng ký và hủy đăng ký một nội dung và đánh giá sự quan tâm của khán giả đối với một kiểu nội dung cụ thể
* Ứng dụng Big Data còn giúp đặt nhóm mục tiêu quảng cáo cho các công ty truyền thông
* Có thể tạo thêm các tính năng mới để phân tích nhu cầu
* Nhà quảng cáo (công ty truyền thông, người nổi tiếng, người phụ trách truyền thông) có thể chọn địa điểm tần xuất phân phối
* Tùy mức độ phổ biến, nghệ sĩ có thể chọn thiết bị, hệ điều hành để phân phối bài hát hoặc video của mình

## Những thách thức trong phân tích dữ liệu lớn

Big Data được đặc trưng bởi tính đa chiều cao và kích thước mẫu lớn [2]. Hai đặc điểm này dẫn đến ba thách thức riêng biệt:

* Thách thức do tính đa chiều cao:
* Tích tụ nhiễu: Với số lượng lớn các biến, lỗi ngẫu nhiên có thể được khuếch đại, dẫn đến kết quả sai lệch.
* Hệ số tương quan giả tạo: Do ngẫu nhiên, một số biến có thể có vẻ liên quan đến nhau khi thực tế không phải vậy. Điều này có thể dẫn đến kết luận sai lầm.
* Tính đồng nhất ngẫu nhiên: Trong các chiều cao, các điểm dữ liệu có thể có xu hướng tập hợp lại với nhau ngay cả khi chúng không có mối quan hệ thực sự, khiến việc xác định các mẫu thực sự trở nên khó khăn.
* Thách thức do tính đa chiều cao và kích thước mẫu lớn:
* Chi phí tính toán cao: Xử lý các tập dữ liệu khổng lồ với nhiều biến đòi hỏi sức mạnh và tài nguyên tính toán đáng kể.
* Sự không ổn định của thuật toán: Một số phương pháp phân tích gặp khó khăn với dữ liệu có chiều cao, dẫn đến kết quả không đáng tin cậy hoặc không thể dự đoán được.
* Thách thức do tính không đồng nhất của dữ liệu:
* Tính không đồng nhất: Dữ liệu từ các nguồn khác nhau có thể có sự không nhất quán về định dạng, phép đo hoặc phương pháp thu thập, khiến việc kết hợp và phân tích hiệu quả trở nên khó khăn.
* Biến đổi thí nghiệm: Sự khác biệt về cách thức thu thập dữ liệu (ví dụ: thời gian, công nghệ) có thể dẫn đến sai lệch làm sai lệch kết quả.
* Sai lệch thống kê: Phương pháp lấy mẫu và kỹ thuật tổng hợp dữ liệu có thể dẫn đến sai lệch cần được giải thích để có được thông tin chính xác.
* Những thách thức này làm nổi bật nhu cầu:
* Quy trình mạnh mẽ hơn: Phát triển các phương pháp phân tích ít bị nhiễu và có thể xử lý dữ liệu có chiều cao hiệu quả.
* Kỹ thuật thích ứng: Tạo các thủ tục có thể điều chỉnh theo các đặc điểm cụ thể của từng tập dữ liệu.

## Các bước cơ bản trong phân tích dữ liệu lớn

### Thu thập dữ liệu

Thu thập dữ liệu là bước đầu tiên giúp doanh nghiệp xác định nguồn tài nguyên dữ liệu sẵn có hiện tại. Một số hoạt động diễn ra trong bước này bao gồm:

Xác định mục tiêu thu thập: Phân tích dữ liệu phục vụ mục đích gì? Bạn cần kết quả nào ở việc phân tích? Bạn chỉ cần liệt kê và trả lời những câu hỏi cần thiết để xác định mục tiêu của mình.

Xác định nguồn gốc và tính chất của dữ liệu: Dữ liệu có thể thu thập từ các hệ thống nào bên trong doanh nghiệp? Hoặc qua các báo cáo hoặc từ các nguồn xác thực nào khác bên ngoài? Định dạng, cấu trúc, và độ chính xác của chúng như thế nào?

Xác định nơi lưu trữ dữ liệu: Dữ liệu nên được lưu trữ ở đâu để an toàn và được sắp xếp có khoa học, đồng thời tiện lợi cho doanh nghiệp truy cập và sử dụng?

### Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là quá trình chuẩn bị dữ liệu để chuẩn bị cho các bước phân tích hoặc mô hình hóa sau này. Một số hoạt động quan trọng có thể kể tới:

Làm sạch dữ liệu: Loại bỏ các dữ liệu sai, không hợp lệ, trùng lặp hoặc bị nhiễu; đồng thời dữ liệu sẽ được kiểm tra, sửa các giá trị lỗi và chuyển đổi định dạng về cùng một đơn vị hoặc thang đo nếu cần.

Xử lý dữ liệu thiếu: Xác định và xử lý các dữ liệu bị thiếu trong tập dữ liệu bằng một số phương pháp như: điền giá trị trung bình, giá trị gần nhất, hoặc dự báo để điền dữ liệu có độ sai số ít nhất. Hoặc trong nhiều trường hợp, các dòng hoặc cột dữ liệu có quá nhiều giá trị thiếu không thể khôi phục sẽ được loại bỏ.

### Khám phá dữ liệu

Khám phá dữ liệu là bước quan trọng để hiểu sâu về dữ liệu trước khi tiến hành phân tích chi tiết hơn, ví dụ như giúp nhận diện các đặc điểm quan trọng, phân loại các đặc tính của dữ liệu, và tạo ra cái nhìn toàn cảnh. Hai hoạt động khám phá cơ bản bao gồm:

Mô tả thống kê: Là hoạt động tổng hợp, mô tả đặc điểm chính của dữ liệu bằng cách sử dụng các thống kê cơ bản như trung bình, phương sai, phân phối và mẫu tổ chức.

Trực quan hóa dữ liệu: Là hoạt động sử dụng biểu đồ và đồ thị để thể hiện sự phân bố, xu hướng và mối quan hệ giữa các biến số trong tập dữ liệu.

### Phân tích dữ liệu

Phân tích dữ liệu là bước cuối cùng và có vai trò quan trọng nhất trong toàn bộ quá trình.

Các hoạt động tại bước này tập trung chủ yếu vào việc sử dụng các công cụ và kỹ thuật phân tích phù hợp để trích xuất thông tin quan trọng từ các dữ liệu đã xử lý, sau đó rút ra những kết luận có ý nghĩa.

Kết quả phân tích có thể làm phát sinh những câu hỏi mới, đòi hỏi việc thu thập thêm dữ liệu hoặc tiến hành các phân tích bổ sung.

# CÂY QUYẾT ĐỊNH

## Khái niệm về cây quyết định

Trong lĩnh vực máy học, cây quyết định là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Mỗi một nút trong (internal node) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định.

Đặc điểm chính của cây quyết định là một tập hợp con đệ quy của trường dữ liệu đích theo giá trị của các trường đầu vào liên quan.

Cây quyết định là một công cụ khá phổ biến trong nhiều dạng ứng dụng, với cơ chế rút trích các luật dạng if – then dựa trên tập dữ liệu mẫu. Sau đây là định nghĩa cây quyết định.

Cây quyết định là dạng đặc biệt của cấu trúc cây nhằm hỗ trợ cho tiến trình ra quyết định. Cây gồm các thành phần:

* Nút trong: biểu diễn thuộc tính
* Nhánh: biểu diễn giá trị cụ thể của thuộc tính đó,
* Nút Lá: đại diên cho giá trị dự đoán của thuộc tính mục tiêu

Đặc điểm chính của cây quyết định bao gồm:

* Phân cấp: Cây quyết định có cấu trúc phân cấp, với mỗi nút cha phân chia dữ liệu thành các tập con dựa trên các điều kiện đặc biệt.
* Quyết định dựa trên điều kiện: Mỗi nút trong cây đại diện cho một điều kiện hoặc một biến và dùng để quyết định việc phân chia dữ liệu.
* Phương pháp tìm kiếm tối ưu: Cây quyết định có thể được xây dựng bằng nhiều phương pháp tìm kiếm tối ưu như CART (Classification and Regression Trees), ID3 (Iterative Dichotomiser 3), C4.5, và Gini Impurity.
* Dễ hiểu và diễn giải: Cây quyết định tạo ra các quy tắc dễ hiểu và diễn giải được, giúp trong việc giải thích quyết định và phân tích dữ liệu.
* Khả năng xử lý dữ liệu không đồng nhất: Cây quyết định có thể xử lý dữ liệu không đồng nhất, bao gồm cả dữ liệu có giá trị thiếu.
* Dễ dàng sử dụng và triển khai: Cây quyết định có thể được dễ dàng triển khai trong các ứng dụng thực tế và thường được sử dụng trong nhiều lĩnh vực như học máy, khoa học dữ liệu, và dự đoán.

Học bằng cây quyết định cũng là một phương pháp thông dụng trong khai phá dữ liệu. Quá trình học trên cây quyết định được dựa trên tập dữ liệu huấn luyện có dạng bảng, ở đó các cột là các thuộc tính điều kiện và một thuộc tính mục tiêu (quyết định), mỗi dòng sẽ được gán nhãn là giá trị của thuộc tính mục tiêu.

Ngoài ra một cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình đệ quy hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại rừng ngẫu nhiên (random forest) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.

## Một số thuật toán tạo cây quyết định

### ID3 (Iterative Dichotomiser 3):

* Là thuật toán tạo cây quyết định đầu tiên và được sử dụng rộng rãi.
* Sử dụng Entropy (Entropy) để đo lường độ thuần nhất của tập dữ liệu và Information Gain (Information Gain) để chọn thuộc tính tốt nhất để phân chia tập dữ liệu.
* Điểm mạnh: Dễ hiểu, dễ triển khai.
* Điểm yếu: Khả năng xử lý dữ liệu nhiễu và dữ liệu thiếu giá trị không tốt.

### C4.5 (C4.5):

* Là phiên bản nâng cấp của ID3, khắc phục một số nhược điểm của ID3.
* Sử dụng Gain Ratio (Gain Ratio) thay cho Information Gain để chọn thuộc tính tốt nhất để phân chia tập dữ liệu.
* C4.5 cũng hỗ trợ xử lý dữ liệu nhiễu và dữ liệu thiếu giá trị.
* Điểm mạnh: Khắc phục nhược điểm của ID3, xử lý dữ liệu nhiễu và dữ liệu thiếu giá trị tốt hơn.
* Điểm yếu: Phức tạp hơn ID3, tốn thời gian tính toán hơn.

### CART (Classification and Regression Trees):

* Sử dụng Gini impurity (Gini impurity) để đo lường độ thuần nhất của tập dữ liệu và sử dụng tiêu chí Gini Index (Gini Index) để chọn thuộc tính tốt nhất để phân chia tập dữ liệu.
* CART có thể tạo ra cả cây quyết định phân loại và hồi quy.
* Điểm mạnh: Có thể tạo ra cả cây quyết định phân loại và hồi quy, xử lý dữ liệu nhiễu và dữ liệu thiếu giá trị tốt.
* Điểm yếu: Phức tạp hơn ID3 và C4.5, tốn thời gian tính toán hơn.

### CHAID (Chi-squared Automatic Interaction Detection):

* Sử dụng thống kê chi bình phương (chi-square) để chọn thuộc tính tốt nhất để phân chia tập dữ liệu.
* CHAID có thể xử lý cả dữ liệu phân loại và dữ liệu số.
* Điểm mạnh: Dễ hiểu, dễ triển khai, xử lý cả dữ liệu phân loại và dữ liệu số tốt.
* Điểm yếu: Khả năng xử lý dữ liệu nhiễu và dữ liệu thiếu giá trị không tốt.

### PRISM (PRuning and Imputation from Samples using Multi-Dimensional Search):

* Sử dụng kết hợp nhiều phương pháp khác nhau để tạo cây quyết định, bao gồm ID3, C4.5 và CHAID.
* PRISM có thể xử lý cả dữ liệu phân loại và dữ liệu số.
* Điểm mạnh: Khả năng xử lý dữ liệu nhiễu và dữ liệu thiếu giá trị tốt, xử lý cả dữ liệu phân loại và dữ liệu số tốt.
* Điểm yếu: Phức tạp hơn các thuật toán khác, tốn thời gian tính toán hơn.

### Một số thuật toán tạo cây quyết định khác:

* MARS (Multivariate Adaptive Regression Splines)
* QUEST (Quest for the Optimal Split)
* Random Forest
* Gradient Boosting Trees

## Đánh giá hiệu quả của cây quyết định

### Ưu điểm

Thuật toán Decision trees , trực quan, không quá phức tạp, khác với các thuật toán ví dụ như Artificial Neural network không thể hiện rõ quy luật phân loại. Đồng thời bộ dữ liệu training không nhất thiết phải quá lớn để tiến hành xây dựng mô hình phân tích.

Một số thuật toán cây quyết định có khả năng xử lý dữ liệu bị missing và dữ liệu bị lỗi mà không cần áp dụng phương đơn giảnpháp như “imputing missing values” hay loại bỏ. Bên cạnh đó Decision trees ít bị ảnh hưởng bởi các dữ liệu ngoại lệ (outliers)

Thuật toán cây quyết định là phương pháp không sử dụng tham số, “nonparametric”, nên không cần phải có các giả định ban đầu về các quy luật phân phối như trong thống kê, và nhờ đó kết quả phân tích có được là khách quan, “tự nhiên” nhất.

Thuật toán cây quyết định có thể giúp chúng ta phân loại đối tượng dữ liệu theo biến mục tiêu có nhiều lớp, nhiều nhóm khác nhau (multi-class classification) đặc biệt nếu biến mục tiêu là dạng biến định lượng phức tạp.

Thuật toán cây quyết định có thể áp dụng linh hoạt cho các biến target, biến mục tiêu là biến định tính (classification task) ví dụ phân loại khách hàng theo “rủi ro tín dụng” và “không rủi ro tín dụng” , và cả định lượng (regression task) ví dụ ước lượng xác suất khách hàng có rủi ro tín dụng

Thuật toán cây quyết định mang lại kết quả dự báo có độ chính xác cao, dễ dàng thực hiện, nhanh chóng trong việc huấn luyện, không cần phải chuyển đổi các biến vì kết quả sẽ như nhau với bất kể loại biến dữ liệu biến đổi ra sao.

Dựa trên quy luật ra quyết định (Decision rule) để xây dựng nên thuật toán cây quyết định rất dễ diễn giải hay giải thích đến người nghe, người xem – những người muốn hiểu rõ về kết quả phân tích nhưng không có kiến thức gì về khoa học dữ liệu.

Thuật toán cây quyết định vẫn nói lên được mối liên hệ giữa các biến, các thuộc tính dữ liệu một cách trực quan nhất mặc dù không thể hiện được rõ mối quan hệ tuyến tính, hay mức độ liên hệ giữa chúng như phương pháp phân tích hồi quy (regression analysis) có được.

Ngoài kinh tế, tài chính, thuật toán cây quyết định có thể được ứng dụng trong lĩnh vực y tế, nông nghiệp, sinh học.

### Nhược điểm

Thuật toán cây quyết định hoạt động hiệu quả trên bộ dữ liệu đơn giản có ít biến dữ liệu liên hệ với nhau, và ngược lại nếu áp dụng cho bộ dữ liệu phức tạp.

Cụ thể, thuật toán cây quyết định khi được áp dụng với bộ dữ liệu phức tạp, nhiều biến và thuộc tính khác nhau có thể dẫn đến mô hình bị overfitting, quá khớp với dữ liệu training dẫn đến vấn đề không đưa ra kết quả phân loại chính xác khi áp dụng cho dữ liệu test, và dữ liệu mới.

Đối với thuật toán cây quyết định khi có sự thay đổi nhỏ trong bộ dữ liệu có thể gây ảnh hưởng đến cấu trúc của mô hình. Nghĩa là khi chúng ta điều chỉnh dữ liệu, cách thức phân nhánh, ngắt cây sẽ bị thay đổi, có thể dẫn đến kết quả sẽ khác so với ban đầu, phức tạp hơn. Các chuyên gia gọi đây là vấn đề “high variance” – giá trị phương sai cao.

Thuật toán cây quyết định chỉ áp dụng cho biến định tính (classification tree) nếu phân loại sai có thể dẫn đến sai lầm nghiêm trọng ví dụ một người có khả năng bị đột quỵ lại được phân loại là không thì vô tính đặt người này vào tình thế nguy hiểm. Còn đối với thuật toán cây quyết định áp dụng cho biến định lượng (regression tree), thì chỉ phân loại đối tượng, hay dự báo theo phạm vi giá trị (range) được tạo ra trước đó, vì vậy đây cũng là một hạn chế khi khả năng có nhiều phạm vi giá trị khác mà thuật toán chưa xét đến.

Thuật toán cây quyết định có khả năng “bias” hay thiên vị nếu bộ dữ liệu không được cân bằng. Nói đơn giản, khi bộ dữ liệu được phân ra thành các nhóm theo các đặc trưng khác nhau nào đó, mà số lượng quan sát trong mỗi nhóm là quá chênh lệch hay khác biệt rõ rệt về đặc trưng, lúc này có thể dẫn đến mô hình bị “bias”, phân nhánh đơn giản, chỉ xét đến các giá trị tiêu biểu, và nguy cơ “Underfitting” (không rà soát hết các khả năng phân loại dữ liệu). Tuy nhiên khi mô hình quá phức tạp, mọi biến dữ liệu đều có khả năng phân nhánh và làm cơ sở phân loại các đối tượng dữ liệu, thì lúc này “bias” ở mức thấp nhưng nguy cơ không thể áp dụng được dữ liệu mới.

Thuật toán cây quyết định yêu cầu bộ dữ liệu training và test phải được chuẩn bị hoàn hảo, chất lượng tốt phải được cân đối theo các lớp, các nhóm trong biến mục tiêu, ví dụ có sự chênh lệch không quá lớn giữa số lượng đối tượng dữ liệu thuộc lớp A của biến mục tiêu và số lượng đối tượng dữ liệu thuộc lớp B của biến mục tiêu. Ngoài ra biến mục tiêu phải có các giá trị “rời rạc” dễ nhận biết, không được quá đa dạng, và phải cụ thể để quá trình phân loại diễn ra dễ dàng hơn cho thuật toán.

Thuật toán cây quyết định được hình thành trên các cách thức phân nhánh tại mỗi một thời điểm bất kỳ, ở một node hay biến dữ liệu bất kỳ và chỉ quan tâm duy nhất vào việc phân nhánh sao cho tối ưu tại thời điểm ấy, chứ không xét đến toàn bộ mô hình phải được thiết lập hiệu quả ra sao. Do đó sẽ có trường hợp bạn cảm thấy việc phân nhánh dễ dàng, cứ thế tiếp tục cho đến khi không còn đối tượng dữ liệu để phân loại nhưng khi kết thúc nhìn mô hình lại quá cồng kềnh, phức tạp. Lúc này không thể tìm ra nguyên nhân. Đây cũng chính là một trong những khuyết điểm của Decision trees.

Thuật toán cây quyết định không “hỗ trợ” kỹ thuật hay khả năng “truy vấn ngược” mà chỉ phân nhánh liên tục dựa trên các công thức phân nhánh cho đến khi thấy được kết quả sau cùng nên chúng ta khó phát hiện được các lỗi ở đâu nếu có sai sót.

## Ứng dụng của cây quyết định

Đánh giá cơ hội tăng trưởng trong tương lai: Một trong những ứng dụng của cây quyết định liên quan đến việc đánh giá các cơ hội tăng trưởng tiềm năng cho doanh nghiệp dựa trên dữ liệu lịch sử. Dữ liệu lịch sử về doanh số bán hàng có thể được sử dụng trong cây quyết định, có thể dẫn đến những thay đổi căn bản trong chiến lược của doanh nghiệp nhằm hỗ trợ việc mở rộng và tăng trưởng.

Sử dụng dữ liệu nhân khẩu học để tìm kiếm khách hàng tiềm năng: Một ứng dụng khác của cây quyết định là sử dụng dữ liệu nhân khẩu học để tìm kiếm khách hàng tiềm năng. Họ có thể giúp hợp lý hóa ngân sách tiếp thị và đưa ra quyết định sáng suốt về thị trường mục tiêu mà doanh nghiệp tập trung vào. Trong trường hợp không có cây quyết định, doanh nghiệp có thể dành thị trường tiếp thị của mình mà không lưu ý đến nhân khẩu học cụ thể, điều này sẽ ảnh hưởng đến tổng doanh thu của doanh nghiệp.

Đóng vai trò là công cụ hỗ trợ trong một số lĩnh vực

Người cho vay cũng sử dụng cây quyết định để dự đoán xác suất khách hàng không trả được nợ nhờ áp dụng việc tạo mô hình dự đoán bằng cách sử dụng dữ liệu trong quá khứ của khách hàng. Việc sử dụng công cụ hỗ trợ cây quyết định có thể giúp người cho vay đánh giá mức độ tín nhiệm của khách hàng để ngăn ngừa tổn thất.

Cây quyết định cũng có thể được sử dụng để nghiên cứu hoạt động trong việc lập kế hoạch hậu cần và quản lý chiến lược. Họ có thể giúp xác định các chiến lược phù hợp sẽ giúp công ty đạt được các mục tiêu đã định. Các lĩnh vực khác có thể áp dụng cây quyết định bao gồm kỹ thuật, giáo dục, luật, kinh doanh, y tế và tài chính, tiếp thị,...

Tài chính: Trong lĩnh vực tài chính, cây quyết định thường được sử dụng trong việc chấm điểm tín dụng, phát hiện gian lận và ra quyết định đầu tư. Bằng cách phân tích các mẫu và xác định các điểm bất thường trong bộ dữ liệu lớn, cây quyết định có thể giúp xác định các hoạt động gian lận tiềm ẩn. Cấu trúc cây cho phép tạo ra các quy tắc và điều kiện phân loại giao dịch là gian lận hoặc hợp pháp, giúp các tổ chức tài chính và doanh nghiệp ngăn chặn các hoạt động gian lận và giảm thiểu tổn thất.

Y tế: Trong lĩnh vực y tế, cây quyết định có thể hỗ trợ chẩn đoán bệnh, dự đoán kết quả của bệnh nhân và xác định các yếu tố rủi ro. Bằng cách phân tích dữ liệu bệnh nhân, chẳng hạn như triệu chứng, tiền sử bệnh và kết quả xét nghiệm, cây quyết định có thể hỗ trợ chẩn đoán các tình trạng bệnh lý khác nhau. Cấu trúc cây cho phép các chuyên gia chăm sóc sức khỏe tuân theo một loạt câu hỏi và tiêu chí để đưa ra chẩn đoán chính xác, có khả năng tiết kiệm thời gian và cải thiện kết quả của bệnh nhân.

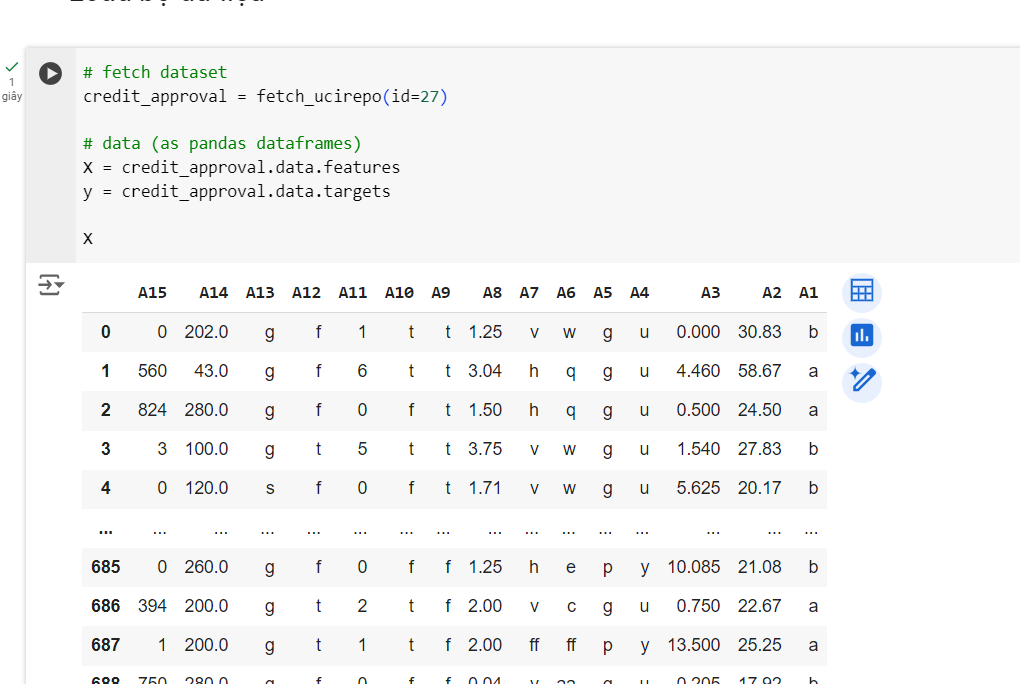
Tiếp thị: Cây quyết định được sử dụng trong tiếp thị để phân khúc khách hàng, dự đoán tỷ lệ rời bỏ và quảng cáo được nhắm mục tiêu. Bằng cách phân tích dữ liệu khách hàng, cây quyết định có thể xác định các nhóm hoặc phân khúc khách hàng khác nhau dựa trên đặc điểm, hành vi hoặc sở thích của họ. Phân khúc này cho phép doanh nghiệp điều chỉnh chiến lược tiếp thị và dịch vụ của mình cho phù hợp với các phân khúc khách hàng cụ thể, cuối cùng là cải thiện sự hài lòng và giữ chân khách hàng.

# MÔ HÌNH PHÊ DUYỆT THẺ TÍN DỤNG

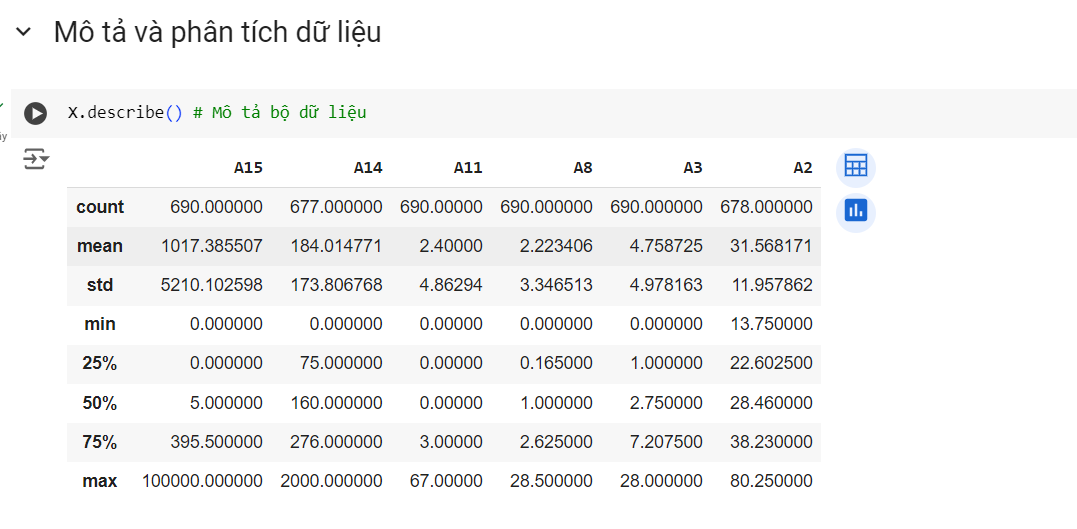
## Giới thiệu

### Tập hợp dữ liệu

Dữ liệu được nhóm tác giả thu thập trên UC Irvine Machine Learning Repository có tên là Credit Approval gồm 690 dòng và 15 cột, có tên các cột hiển thị như hình dưới đây:



Hình 1: Tập hợp dữ liệu



Hình 2: Mô tả và phân tích dữ liệu

Bảng này cung cấp các thông số thống kê như số lượng, trung bình, độ lệch chuẩn, giá trị nhỏ nhất, các phân vị 25%, 50%, 75% và giá trị lớn nhất cho các biến số A15, A14, A11, A8, A3 và A2.

### Làm sạch dữ liệu

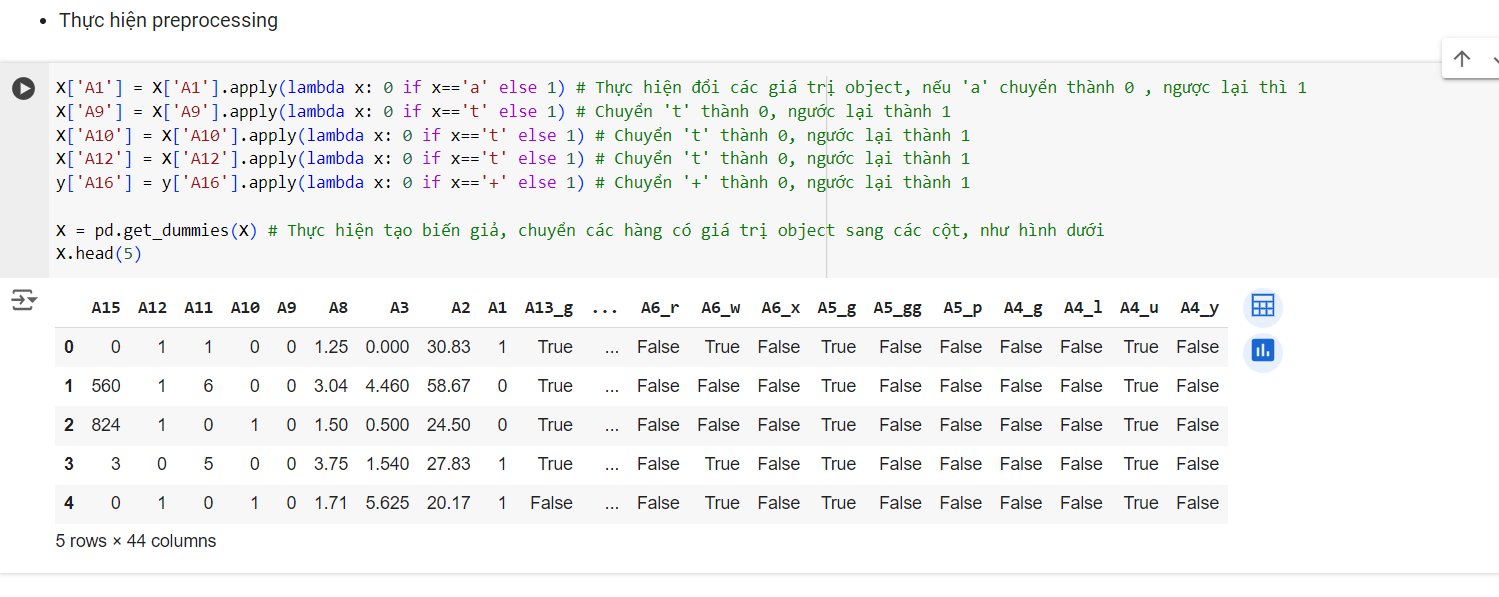


Hình 3: Xử lý dữ liệu

* Xóa cột A14 để cột này được xử lý như zip code.
* Chuyển đổi cột A2 từ kiểu object sang kiểu số.
* Thay thế các giá trị '?' thành Null (NaN).
* Hiển thị 5 hàng đầu của DataFrame X sau khi thực hiện các thay đổi trên.

Bảng này hiển thị các giá trị trong 14 cột khác nhau, từ A15 đến A1, cho 5 dòng đầu tiên. Các giá trị ở đây bao gồm các số nguyên, số thực và ký tự.

Tiếp theo ta gán các giá trị còn thiếu bằng giá trị trung bình của từng cột, nhận thấy được cột A1 vẫn còn giá trị thiếu vì có kiểu dữ liệu là Object nên ta tiếp tục tạo vòng lặp để gán giá trị còn thiếu vào đó. Cho đến đây, dữ liệu vẫn còn một số cột là kiểu chữ, ta thực hiện chuyển chữ thành số để chuẩn hoá dữ liệu.



Hình 4: Chuyển các ký tự chữ thành số

