**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**MÔN KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:**

**SỬ DỤNG PHƯƠNG PHÁP PHÂN CỤM KMEANS CLUSTERING PHÂN LOẠI THÀNH PHẦN CỦA RƯỢU VANG**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Sinh viên thực hiện** | **: NGUYỄN MẠNH CƯỜNG**  **: PHẠM ĐỨC HÙNG**  **: NGUYỄN MẠNH THỜI** | | **Giảng viên hướng dẫn** | **: VŨ VĂN ĐỊNH** | | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | | **Lớp** | **: D14CNPM1** | | | **Khóa** | **: 2019-2024** | | |  |
| ***Hà Nội, tháng 5 năm 2022*** |  |

**PHIẾU CHẤM ĐIỂM**

**Sinh viên thực hiện**

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi Chú** |
| Nguyễn Mạnh Cường |  |  |
| Phạm Đức Hùng |  |  |
| Nguyễn Mạnh Thời |  |  |

**Giảng viên chấm**

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1 |  |  |
| Giảng viên chấm 2 |  |  |

**MỤC LỤC**

[**CHƯƠNG 1. GIỚI THIỆU VỀ KHAI PHÁ DỮ LIỆU** 2](#_Toc105488201)

[**1.1. Đặt vấn đề** 2](#_Toc105488202)

[**1.2. Khai phá dữ liệu là gì ? 3**](#_Toc105488203)

[**1.3. Các bước để khám phá tri thức 3**](#_Toc105488204)

[**1.4. Các phương pháp chính trong khai phá dữ liệu 4**](#_Toc105488205)

[1.4.1. Phương pháp luật kết hợp 4](#_Toc105488206)

[1.4.2. Phương pháp cây quyết định 5](#_Toc105488208)

[1.4.3. Phương pháp K-Mean 6](#_Toc105488209)

[**1.5. Các công cụ khai phá dữ liệu** 6](#_Toc105488210)

[1.5.1. RapidMiner 6](#_Toc105488211)

[1.5.2. Weka 6](#_Toc105488212)

[1.5.3. KNime 6](#_Toc105488213)

[1.5.4. Apache Mahout 6](#_Toc105488214)

[1.5.5. Oracle Data Mining 6](#_Toc105488215)

[**1.6. Ứng dụng của khai phá dữ liệu** 7](#_Toc105488216)

[1.6.1.Lĩnh vực tài chính 7](#_Toc105488217)

[1.6.2. Lĩnh vực chăm sóc sức khỏe 7](#_Toc105488218)

[1.6.3. Lĩnh vực viễn thông 7](#_Toc105488219)

[1.6.4. Lĩnh vực Marketing và Sales 8](#_Toc105488220)

[1.6.5. Lĩnh vực thương mại điện tử 8](#_Toc105488221)

[1.6.6. Lĩnh vực giáo dục 8](#_Toc105488222)

[1.6.7. Lĩnh vực kỹ thuật sản xuất 8](#_Toc105488223)

[1.6.8. Các ứng dụng khoa học khác 9](#_Toc105488224)

[**CHƯƠNG 2: THUẬT TOÁN K-MEANS CLUSTERING** 10](#_Toc105488225)

[**2.1. Giới thiệu về K-Means Clustering** 10](#_Toc105488226)

[**2.2. Thuật toán K-Means Clustering là gì ?** 10](#_Toc105488227)

[**2.3. Ý tưởng của thuật toán K-Means Clustering** 11](#_Toc105488228)

[**2.4. Ưu điểm và nhược điểm của thuật toán K-Means Clustering** 11](#_Toc105488229)

[**2.5.Các ứng dụng của K-Means Clustering** 12](#_Toc105488232)

[**CHƯƠNG 3: SỬ DỤNG PHƯƠNG PHÁP PHÂN CỤM K-MEANS CLUSTERING PHÂN LOẠI THÀNH PHẦN RƯỢU VANG** 13](#_Toc105488233)

[**3.1. Dữ liệu Wine Dataset for Clustering:** 13](#_Toc105488234)

[3.1.1. Phát biểu bài toán: 13](#_Toc105488235)

[3.1.2. Bộ dữ liệu: 13](#_Toc105488236)

[**3.2. Xử lý dữ liệu** 14](#_Toc105488237)

[**3.3. Kết quả** 15](#_Toc105488238)

[3.3.1. Đọc dữ liệu 15](#_Toc105488239)

[3.3.2. Đồ thị 2D 15](#_Toc105488240)

# **LỜI MỞ ĐẦU**

Những năm gần đây, các hệ thống cơ sở dữ liệu đã đem lại những lợi ích vô cùng to lớn cho con người. Song hành cùng sự phát triển nhanh chóng của công nghệ thông tin và những ứng dụng của nó trong đời sống, kinh tế và xã hội, lượng dữ liệu thu thập ngày càng nhiều theo thời gian, dẫn đến việc xuất hiện ngày càng nhiều các hệ thống cơ sở dữ liệu có kích thước lớn. Trong xã hội hiện đại, thông tin được coi như sức mạnh và là yếu tố quyết định thành công trong mọi lĩnh vực, do đó việc tìm ra thông tin hữu ích trong khối dữ liệu khổng lồ được xem như mục tiêu hàng đầu của mọi tổ chức và cá nhân.

Với tầm quan trọng của dữ liệu lớn trong mọi mặt của đời sống xã hội như thế, các ngành học liên quan đến kỹ thuật xử lý các dữ liệu này như công nghệ thông tin, khoa học dữ liệu, khoa học máy tính... đang ngày càng trở nên hấp dẫn bởi sự gia tăng nhu cầu lao động chất lượng cao của xã hội. Một trong các kỹ năng quan trọng mà các chuyên ngành này đều phải học qua, đó là data mining (tiếng Việt là khai phá dữ liệu). Đặc biệt, data mining không chỉ phổ biến đối với những người học về công nghệ mà còn ứng dụng trong rất nhiều ngành nghề, lĩnh vực như marketing, quản lý chuỗi cung ứng - logistic, kinh doanh, sản xuất... Theo đó, data mining là quá trình đi sâu vào bộ dữ liệu để phân tích và tìm kiếm các chi tiết, giá trị ẩn bên trong từng dữ liệu hay muốn xem dữ liệu cung cấp thông tin gì, có ích hay không. Data mining được xem là lợi thế của các công ty hoạt động trong ngành sản xuất, marketing...

Chính bởi tính ứng dụng rộng rãi mà Khai phá dữ liệu mang lại, vậy nên trong quá trình học và nghiên cứu tại trường Đại học Điện Lực khi học và tìm hiểu bộ môn **“ Khai phá dữ liệu ”** của thầy **“ Vũ Văn Định ”** , nhóm chúng em đã chọn đề tài **“ Sử dụng phương pháp phân cụm Kmeans Clustering phân loại thành phần của rượu vang ”** làm đề tài nghiên cứu để tìm hiểu sâu hơn về Machine Learning.

# **CHƯƠNG 1. GIỚI THIỆU VỀ KHAI PHÁ DỮ LIỆU**

## **1.1. Đặt vấn đề**

Thuật ngữ Khai phá dữ liệu ra đời vào cuối những năm 80 thế kỷ trước. Có nhiều định nghĩa khác nhau về Khai phá dữ liệu, nhưng diễn đạt một cách dễ hiểu thì Khai phá dữ liệu là quá trình tìm kiếm những thông tin (tri thức) có ích, tiềm ẩn và mang tính dự đoán trong các khối cơ sở dữ liệu lớn.

Mục đích việc phát hiện tri thức từ dữ liệu Khai phá dữ liệu là cốt lõi của **quá trình khám phá tri thức** gồm có các giải thuật chuyên dùng, dưới một số quy định về hiệu quả tính toán chấp nhận được. Khai phá dữ liệu nhằm tìm ra những mẫu mới, những thông tin tiềm ẩn mang tính dự đoán chưa được biết đến, có khả năng mang lại lợi ích cho người sử dụng và tìm ra các mẫu được quan tâm nhất tồn tại trong CSDL, nhưng chúng lại bị che giấu bởi một số lượng lớn dữ liệu.

Ngày nay, công nghệ thông tin phát triển đồng nghĩa với việc phát triển các phần mềm ứng dụng. Phần mềm **Khai phá dữ liệu** là một công cụ phân tích dùng để phân tích dữ liệu. Nó cho phép người sử dụng phân tích dữ liệu theo nhiều góc nhìn khác nhau, phân loại dữ liệu theo những quan điểm riêng biệt và tổng kết các mối quan hệ đã được bóc tách.

Hiện kỹ thuật **Khai phá dữ liệu** đang được áp dụng một cách rộng rãi trong nhiều lĩnh vực kinh doanh và đời sống khác nhau, như: Thương mại (phân tích dữ liệu bán hàng và thị trường, phân tích đầu tư, quyết định cho vay, phát hiện gian lận,...); Thông tin sản xuất (điều khiển và lập kế hoạch, hệ thống quản lý, phân tích kết quả thử nghiệm…); Thông tin khoa học (dự báo thời tiết,...); CSDL sinh học (ngân hàng gen,…); Khoa học địa lý (dự báo động đất,…); Trong y tế, marketing, ngân hàng, viễn thông, du lịch, internet…

Những gì thu được từ **Khai phá dữ liệu** thật đáng giá. Điều đó được chứng minh bằng thực tế: Chẩn đoán bệnh trong y tế dựa trên kết quả xét nghiệm đã giúp cho bảo hiểm y tế phát hiện ra nhiều trường hợp xét nghiệm không hợp lý; Trong dịch vụ viễn thông đã phát hiện ra những nhóm người thường xuyên gọi cho nhau bằng mobile và thu lợi hàng triệu USD; IBM Suft-Aid đã áp dụng **Khai phá dữ liệu** vào phân tích các lần đăng nhập Web vào các trang liên quan đến thị trường để phát hiện sở thích khách hàng, từ đó đánh giá hiệu quả của việc tiếp thị qua Web và cải thiện hoạt động của các Website; trang Web mua bán qua mạng cũng tăng doanh thu nhờ áp dụng **Khai phá dữ liệu** trong việc phân tích sở thích mua bán của khách hàng.

## **1.2. Khai phá dữ liệu là gì ?**

Khai phá dữ liệu (Data mining – DM) – Khám phá tri thức từ dữ liệu (Knowledge discovery from data) là việc trích rút ra được các mẫu hoặc tri thức quan trọng từ một lượng dữ liệu (rất) lớn.

**Các tên gọi khác:**

- Khám phá tri thức trong các cơ sở dữ liệu (Knowledge discovery

in databases in databases - KDD)

- Trích rút tri thức (Knowledge extraction)

- Phân tích mẫu/dữ liệu (Data/pattern analysis)

- …

**Khai phá dữ liệu khác với :**

- Tìm kiếm thông tin (Information retrieval)

- Xử lý các câu truy vấn (SQL) đối với các cơ sở dữ liệu

## **1.3. Các bước để khám phá tri thức**

Bước 1- Trích chọn dữ liệu (data selection): Là bước trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases,data warehouses).

Bước 2 - Tiền xử lý dữ liệu (data preprocessing): Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán,...), rút gọn dữ liệu (sử dụng các phương pháp thu gọn dữ liệu, histograms, lấy mẫu,...), rời rạc hoá dữ liệu (dựa vào histograms, entropy, phân khoảng,…). Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và được rời rạc hoá.

Bước 3- Biến đổi dữ liệu (data transformation): Là bước chuẩn hoá và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ cho các kỹ thuật khai phá ở bước sau.

Bước 4- Khai phá dữ liệu (data mining): Đây là bước quan trọng và tốn nhiều thời gian nhất của quá trình **khám phá tri thức**, áp dụng các kỹ thuật khai phá (phần lớn là các kỹ thuật của machine learning) để khai phá, trích chọn được các mẫu (pattern) thông tin, các mối liên hệ đặc biệt trong dữ liệu.

Bước 5- Đánh giá và biểu diễn tri thức (knowledge representation & evaluation): Dùng các kỹ thuật hiển thị dữ liệu để trình bày các mẫu thông tin (tri thức) và mối liên hệ đặc biệt trong dữ liệu đã được khai phá ở bước trên biểu diễn theo dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật,… Đồng thời, bước này cũng đánh giá những tri thức khai phá được theo những tiêu chí nhất định.

Trong giai đoạn KPDL, có thể cần sự tương tác của người dùng để điều chỉnh và rút ra các tri thức cần thiết nhất. Các**tri thức** nhận được cũng có thể được lưu và sử dụng lại.

## **1.4. Các phương pháp chính trong khai phá dữ liệu**

### **1.4.1. Phương pháp luật kết hợp**

Khai phá luật kết hợp là một kĩ thuật quan trọng của khai phá dữ liệu. Mục tiêu nhằm phát hiện mối quan hệ giữa các mục dữ liệu trong cơ sở dữ liệu.

Mô hình đầu tiên của bài toán KPLKH là mô hình nhị phân (hay còn gọi là mô hình cơ bản) được R. Agrawal, T. Imielinski và A. Swami đề xuất vào năm 1993, xuất phát từ nhu cầu phân tích dữ liệu của cơ sở dữ liệu giao tác, phát hiện các mối quan hệ giữa các tập mục hàng hóa (Itemsets) đã bán được tại các siêu thị.

Việc xác định các quan hệ này không phân biệt vai trò khác nhau cũng như không dựa vào các đặc tính dữ liệu vốn có của các mục dữ liệu mà chỉ dựa vào sự xuất hiện cùng lúc của chúng.

**Bài toán khai phá luật kết hợp:**

Bài toán KPLKH có thể phát biểu như sau: Cho cơ sở dữ liệu giao tác DB, ngưỡng độ hỗ trợ tối thiểu minsup và ngưỡng độ tin cậy tối thiểu minconf.

Yêu cầu: Tìm tất cả các luật kết hợp X→Y trên cơ sở dữ liệu DB sao cho sup(X→Y) > minsup và conf (X→Y) > minconf.

KPLKH này được gọi là bài toán cơ bản hay bài toán nhị phân, vì ở đây, giá trị của mục dữ liệu trong cơ sở dữ liệu là 0 hoặc 1 (xuất hiện hay không xuất hiện).

Bài toán KPLKH trong cơ sở dữ liệu chia thành hai bài toán con:

(1) Tìm tất cả các tập mục thường xuyên: Một tập mục là thường xuyên được xác định qua tính độ hỗ trợ và thoả mãn độ hỗ trợ cực tiểu.

(2) Sinh ra các luật kết hợp từ các tập mục thường xuyên đã tìm được thỏa mãn độ tin cậy tối thiểu cho trước.

Khi KPLKH trong cơ sở dữ liệu DB thì mọi khó khăn nằm ở bài toán thứ nhất là tìm tập mục thường xuyên.

Thuật ngữ [Khai phá dữ liệu](https://vietnambiz.vn/khai-pha-du-lieu-data-mining-la-gi-nhung-dac-diem-can-luu-y-20191130175442498.htm) ra đời vào cuối những năm 80 thế kỷ trước. Có nhiều định nghĩa khác nhau về khai phá dữ liệu, nhưng diễn đạt một cách dễ hiểu thì khai phá dữ liệu là quá trình tìm kiếm những thông tin (tri thức) có ích, tiềm ẩn và mang tính dự đoán trong các khối cơ sở dữ liệu lớn.

### **1.4.2. Phương pháp cây quyết định**

Cây quyết định cũng là một phương pháp thông dụng trong [khai phá dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u). Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó. Một cây quyết định có thể được học bằng cách chia [tập hợp](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_h%E1%BB%A3p) 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](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%87_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](https://vi.wikipedia.org/w/index.php?title=R%E1%BB%ABng_ng%E1%BA%ABu_nhi%C3%AAn&action=edit&redlink=1) (*random forest*) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.

Cây quyết định còn có hai tên khác:

**Cây hồi quy** (*Regression tree*) ước lượng các hàm giá có giá trị là [số thực](https://vi.wikipedia.org/wiki/S%E1%BB%91_th%E1%BB%B1c) thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)

**Cây phân loại** (*Classification tree*), nếu *y* là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

Diagram

Description automatically generated

*Mô hình cây quyết định*

### **1.4.3. Phương pháp K-Mean**

Có nhiều phương pháp được sử dụng trong phân cụm, phương pháp k-Mean được coi là các kỹ thuật cơ bản của phân cụm. Với phương pháp này sẽ chia tập có n đối tượng thành k cụm sao cho các đối tượng trong cùng một cụm thì giống nhau, các đối tượng khác cụm thì khác nhau.

## **1.5. Các công cụ khai phá dữ liệu**

### **1.5.1. RapidMiner**

Công cụ đầu tiên phải kể tới đó là RapidMiner. Đây là công cụ khai phá dữ liệu khá phổ biến hiện nay. Được viết trên nền tảng JAVA nhưng không yêu cầu mã hóa để vận hành.

Ngoài ra, nó còn cung cấp các chức năng khai thác dữ liệu khác nhau như tiền xử lý dữ liệu, biểu diễn dữ liệu, lọc, phân cụm,...

### **1.5.2. Weka**

Công cụ được cho ra đời tại Đại học Wichita là một phần mềm khai thác dữ liệu mã nguồn mở. Tương tự như RapidMiner, công cụ này không yêu cầu mã hóa và sử dụngGUI đơn giản.

Sử dụng Weka, người dùng có thể gọi trực tiếp các thuật toán học máy hoặc nhập chúng bằng mã Java. Weka được trang bị đa dạng chức năng như trực quan hóa, tiền xử lý, phân loại, phân cụm,...

### **1.5.3. KNime**

Với khả năng hoạt động vô cùng mạnh mẽ tích hợp nhiều thành phần khác nhau của học máy và khai phá dữ liệu để cung cấp một nền tảng. KNime hỗ trợ người dùng rất nhiều trong việc xử lý và phân tích dữ liệu, trích xuất, chuyển đổi và tải dữ liệu.

### **1.5.4. Apache Mahout**

Từ nền tảng Big Data Hadoop, người ta đã cho cho ra đời thêm Apache Mahout với mục đích giải quyết nhu cầu ngày càng tăng về khai phá dữ liệu và hoạt động phân tích trong Hadoop. Nó được trang bị nhiều chức năng học máy khác nhau như phân loại, hồi quy, phân cụm,...

### **1.5.5. Oracle Data Mining**

Khi sử dụng Oracle Data Mining. nó cho phép người dùng thực hiện khai phá dữ liệu trên cơ sở dữ liệu SQL để trích xuất các khung hình và biểu đồ. Các phân tích sẽ hiển thị một cách trực quan giúp người dùng dễ dàng đưa ra dự đoán cho kế hoạch tương lai.

## **1.6. Ứng dụng của khai phá dữ liệu**

### **1.6.1.Lĩnh vực tài chính**

Ứng dụng của Data Mining trong lĩnh vực này được dùng để tăng độ trung thành của khách hàng bằng cách thu thập và phân tích dữ liệu hành vi của khách hàng. Thoạt nghe sẽ thấy hơi vô lý vì dữ liệu hành vi người dùng thường sẽ liên quan đến lĩnh vực Marketing.

Tuy nhiên, lĩnh vực tài chính ở đây là các ngân hàng, họ sử dụng nó để dự đoán hành vi của khách hàng để tung ra các dịch vụ và sản phẩm thích hợp.

Ngoài ra, ứng dụng của Data Mining trong lĩnh vực tài chính còn giúp khám phá mối tương quan ẩn giữa các chỉ số tài chính khác nhau để phát hiện ra các hoạt động đáng ngờ có rủi ro tiềm ẩn cao.

Nó còn giúp hỗ trợ xác định các hành động gian lận hoặc không gian lận bằng cách thu thập dữ liệu lịch sử và sau đó biến đổi nó thành thông tin hợp lệ và hữu ích.

### **1.6.2. Lĩnh vực chăm sóc sức khỏe**

Data Mining giúp các doanh nghiệp gia tăng lợi thế cạnh tranh bằng việc cung cấp các thông tin về chính phủ, quy định và đối thủ cạnh tranh. Nó hỗ trợ quá trình R&D và sau đó là chiến lược tiếp cận thị trường với khả năng tiếp cận thông tin nhanh chóng ở mọi giai đoạn.

Thêm vào đó, việc tìm ra mối quan hệ giữa các loại bệnh và hiệu quả của phương pháp điều trị sẽ giúp thay đổi các loại thuốc mới hoặc đảm bảo rằng bệnh nhân được chăm sóc phù hợp, kịp thời.

Ngoài ra, ứng dụng của Data Mining trong lĩnh vực chăm sóc sức khỏe còn hỗ trợ các công ty cung cấp dịch vụ bảo hiểm sức khỏe phát hiện gian lận hay lạm dụng.

### **1.6.3. Lĩnh vực viễn thông**

Trong lĩnh vực này, ứng dụng của data mining giúp ngành viễn thông đạt được lợi thế cạnh tranh và giảm chi phí khách hàng bằng cách hiểu các đặc điểm nhân khẩu học và dự đoán hành vi của khách hàng.

Ngoài ra, nó không những giúp tăng lòng trung thành của khách hàng và cải thiện lợi nhuận bằng cách cung cấp các dịch vụ tùy chỉnh, mà còn hỗ trợ các chiến lược tiếp cận khách hàng bằng cách phát triển các chiến dịch marketing và định giá phù hợp.

### **1.6.4. Lĩnh vực Marketing và Sales**

Về cơ bản, khai phá dữ liệu cho phép các doanh nghiệp hiểu các điều ẩn đằng sau dữ liệu giao dịch mua bán của khách hàng. Từ đó, doanh nghiệp có thể lên kế hoạch và khởi động các chiến dịch marketing mới.

Ứng dụng của Data Mining giúp doanh nghiệp phân tích nhu cầu thị trường để hiểu về sản phẩm thường được mua cùng nhau. Thông tin này giúp doanh nghiệp quảng bá sản phẩm có lợi nhuận cao nhất và tối đa hóa lợi nhuận. Ngoài ra, nó còn khuyến khích khách hàng mua các sản phẩm liên quan.

### **1.6.5. Lĩnh vực thương mại điện tử**

Nhiều công ty thương mại điện tử đang áp dụng ứng dụng của Data Mining để bán hàng qua nhiều nước thông qua các trang web của họ. Một trong những công ty nổi tiếng nhất ứng dụng điều này là Amazon. Họ sử dụng các kỹ thuật khai phá dữ liệu để lái “những người đã xem sản phẩm đó cũng thích sản phẩm được giới thiệu này”.

### **1.6.6. Lĩnh vực giáo dục**

Gần đây có một lĩnh vực mới nổi, nó được gọi là khai phá dữ liệu giáo dục (EDM). Nó khám phá kiến ​​thức từ dữ liệu bắt nguồn từ môi trường giáo dục. Mục tiêu của EDM được xác định là dự đoán hành vi học tập của học sinh trong tương lai.

Chúng tôi sử dụng khai thác dữ liệu để đưa ra quyết định chính xác và cũng như để dự đoán kết quả của học sinh. Với kết quả đó, nhà trường có thể tập trung vào những gì cần dạy và cách dạy từ đó cải tiến các phương pháp dạy học hiệu quả hơn.

Tuy nhiên, ứng dụng của Data Mining trong lĩnh vực giáo dục vẫn chưa thực sự được phát huy mạnh mẽ trong thời điểm hiện tại. Cõ lẽ bởi vì sự mới mẻ của Data Mining ở Việt Nam, cũng như những chi phí nhất định cho việc áp dụng kỹ thuật này vào hệ thống giáo dục.

### **1.6.7. Lĩnh vực kỹ thuật sản xuất**

Chúng ta thường sử dụng các công cụ Data Mining để khám phá các mẫu (patterns) trong quy trình sản xuất phức tạp. Ngoài ra, nó còn được dùng để dự đoán thời gian phát triển sản phẩm, chi phí.

### **1.6.8. Các ứng dụng khoa học khác**

Các ứng dụng của Data Mining được thảo luận ở trên có xu hướng xử lý các tập dữ liệu nhỏ và đồng nhất. Đối với các ứng dụng này thường yêu cầu những kỹ thuật thống kê phù hợp, thu thập một lượng lớn dữ liệu từ các lĩnh vực khoa học, tạo ra một số lượng lớn các tập dữ liệu, v.v.

Sau đây là các ứng dụng khai thác dữ liệu trong lĩnh vực khoa học ứng dụng:

+ Kho dữ liệu và tiền xử lý dữ liệu (Data Warehouses and data preprocessing).

+ Khai thác dựa trên đồ thị (Graph-based mining).

+ Hình dung và kiến ​​thức cụ thể miền (Visualization and domain-specific knowledge).

# **CHƯƠNG 2: THUẬT TOÁN K-MEANS CLUSTERING**

## **2.1. Giới thiệu về K-Means Clustering**

K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát(tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.

Thuật toán phân cụm k-means thường được sử dụng trong các ứng dụng cỗ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu,…

**Chart, scatter chart

Description automatically generated**

*Hình 2.1. Kết quả phân cụm của thuật toán Kmeans*

**2.2. Thuật toán K-Means Clustering là gì ?**

Thuật toán phân cụm k-means là một phương pháp được sử dụng trong phân tích tính chất cụm của dữ liệu. Nó đặc biệt được sử dụng nhiều trong khai phá dữ liệu và thống kê. Nó phân vùng dữ liệu thành k cụm khác nhau. Giải thuật này giúp chúng ta xác định được dữ liệu của chúng ta nó thực sử thuộc về nhóm nào.

Để các bạn dễ hình dung ứng dụng của thuật toán. Chúng ta hãy quan sát một ví dụ thực tế như sau:

Trong các mô hình kinh doanh, doanh nghiệp sẽ chia nhỏ tệp khách hàng ra thành những nhóm đối tượng khác nhau để có thể áp dụng những chiến lược kinh doanh cụ thể cho từng nhóm đối tượng. Điều này giúp cho khách hàng được tiếp cận với các sản phẩm thật sự phù hợp với bản thân họ. Sự phù hợp đó sẽ kéo doanh số của chúng ta tăng lên. Vấn đề đặt ra là làm sao có thể chia nhỏ tệp khách hàng đó ra khi mà số lượng hóa đơn là rất lớn và chúng ta không thể ngồi để phân tích từng vị khách.

Và mục tiêu của các thuật toán phân cụm là từ tập dữ liệu khổng lồ đó. Làm sao chúng ta biết có những nhóm dữ liệu đặc trưng nào trong đó? Từng dữ liệu trong đó thuộc vào nhóm nào? Đó là cái mà thuật toán phân cụm của chúng ta cần đi tìm câu trả lời.

**2.3. Ý tưởng của thuật toán K-Means Clustering**

1. Khởi tạo K điểm dữ liệu trong bộ dữ liệu và tạm thời coi nó là tâm của các cụm dữ liệu của chúng ta.
2. Với mỗi điểm dữ liệu trong bộ dữ liệu, tâm cụm của nó sẽ được xác định là 1 trong K tâm cụm gần nó nhất.
3. Sau khi tất cả các điểm dữ liệu đã có tâm, tính toán lại vị trí của tâm cụm để đảm bảo tâm của cụm nằm ở chính giữa cụm.
4. Bước 2 và bước 3 sẽ được lặp đi lặp lại cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

**2.4. Ưu điểm và nhược điểm của thuật toán K-Means Clustering**

**- Ưu điểm**

Sau đây là một số ưu điểm của thuật toán phân cụm K-Means:

* Nó rất dễ hiểu và dễ thực hiện.
* Nếu chúng ta có số lượng biến lớn thì K-mean sẽ nhanh hơn so với phân cụm phân cấp.
* Khi tính toán lại các centroid, một thể hiện có thể thay đổi cụm.
* Các cụm chặt chẽ hơn được hình thành với K-means so với phân cụm theo thứ bậc.

**- Nhược điểm**

Sau đây là một số nhược điểm của thuật toán phân cụm K-Means:

* Có một chút khó khăn để dự đoán số lượng cụm tức là giá trị của k.
* Đầu ra bị tác động mạnh bởi các đầu vào ban đầu như số cụm (giá trị của k).
* Thứ tự của dữ liệu sẽ có tác động mạnh mẽ đến kết quả cuối cùng.
* Nó rất nhạy cảm với việc thay đổi tỷ lệ. Nếu chúng tôi bán lại dữ liệu của mình bằng phương pháp chuẩn hóa hoặc chuẩn hóa, thì đầu ra sẽ hoàn toàn thay đổi. Đầu ra cuối cùng.
* Sẽ không tốt khi thực hiện công việc phân cụm nếu các cụm có dạng hình học phức tạp.

## **2.5.Các ứng dụng của K-Means Clustering**

- Các mục tiêu chính của phân tích cụm là :

* Để có được trực giác có ý nghĩa từ dữ liệu chúng tôi đang làm việc.
* Cluster-then-dự đoán nơi các mô hình khác nhau sẽ được xây dựng cho các nhóm con khác nhau.

- Để hoàn thành các mục tiêu đề cập ở trên, K-means clustering đang hoạt động đủ tốt. Nó có thể được sử dụng trong các ứng dụng sau:

* Phân khúc thị trường
* Phân cụm tài liệu
* Phân đoạn hình ảnh
* Nén hình ảnh
* Phân khúc khách hàng
* Phân tích xu hướng trên dữ liệu động

# **CHƯƠNG 3: SỬ DỤNG PHƯƠNG PHÁP PHÂN CỤM K-MEANS CLUSTERING PHÂN LOẠI THÀNH PHẦN RƯỢU VANG**

## **3.1. Dữ liệu Wine Dataset for Clustering:**

### **3.1.1. Phát biểu bài toán:**

Bài toán phân loại thành phần của rượu vang dựa trên các thành phần được tìm thấy trong ba loại rượu để phân loại chúng.

+ Giá trị input: Thành phần của rượu

+ Giá trị output: Cụm mà chúng được phân

### **3.1.2. Bộ dữ liệu:**

- Dữ liệu bao gồm các thành phần hóa học được tìm thấy trong ba loại rượu vang :

+ Alcohol : rượu

+ Malic\_Acid : một**loại acid tự nhiên** có vị chua của táo.

+ Ash: các chất vô cơ có trong thực phẩm: khoáng chất,..

+ Ash\_Alcanity: độ kiềm của nước

+ Magnesium: khoáng chất cần thiết cho hệ thần kinh, bảo vệ tim mạch, phát triển xương,…

+ Total\_Phenols: hợp chất hữu cơ tương tự như rượu có tính axit khá cao.

+ Flavanoids: chất chuyển hóa trung gian của thực vật thường gặp trong các loại rau quả.

+ Nonflavanoid\_Phenols: hợp chất phenol có hoạt tính sinh học và một tập hợp con của các chất dinh dưỡng thực vật được tìm thấy nhiều trong các loại thực phẩm thực vật khác nhau.

+ Proanthocyanins: hợp chất có trong vỏ và hạt nho rất giàu chất chống oxy hóa.

+ Color\_Intensity: cường độ màu sắc rượu vang

+ Hue: màu sắc của rượu vang

+ OD280: tỷ lệ tinh khiết protein trong rượu.

+ Proline: loại axit amin thiết yếu có trong rượu vang và nho.

## **3.2. Xử lý dữ liệu**

**Code:**

**Import thư viện:**

**Text

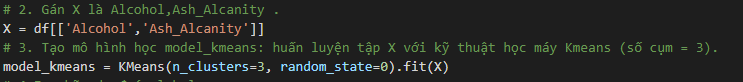
Description automatically generated**

**Đọc file:**

**Text

Description automatically generated**

**Tạo mô hình học model\_kmeans**

****

**In nhãn dự đoán, vẽ tâm cụm**

**A screenshot of a computer

Description automatically generated with medium confidence**

## **3.3. Kết quả**

### **3.3.1. Đọc dữ liệu**

**A screen shot of a computer

Description automatically generated with low confidence**

*Hình 3.1.Kết quả đọc dữ liệu*

### **3.3.2. Đồ thị 2D**

Chart, scatter chart

Description automatically generated

*Hình 3.2. Dữ liệu sau khi phân cụm trên đồ thị 2D*