TRƯỜNG ĐẠI HỌC ĐIỆN LỰC
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN
KHAI PHÁ DỮ LIỆU

ÁP DỤNG THUẬT TOÁN K-MEAN ĐỂ PHÂN CỤM KHÁCH HÀNG CỦA SIÊU THỊ

Sinh viên thực hiện : **VŨ THỊ HOÀI THU**
LÊ THANH BÌNH
NGUYỄN XUÂN THỦY

Giảng viên hướng dẫn : **TS. VŨ VĂN ĐỊNH**

Ngành : **CÔNG NGHỆ THÔNG TIN**

Chuyên ngành : **CÔNG NGHỆ PHẦN MỀM**

Lớp : **D16CNPM7**

Khóa : **2021**

Hà Nội, tháng 5 năm 2024

PHIẾU CHẤM ĐIỂM

STT	Họ và tên sinh viên	Chữ ký	Điểm
1	Vũ Thị Hoài Thu (21810310193)		
2	Nguyễn Xuân Thủy (21810310349)		
3	Lê Thanh Bình (21810310356)		

Họ và tên giảng viên	Chữ ký	Ghi chú
Giảng viên chấm 1:		
Giảng viên chấm 2:		

MỤC LỤC

DANH MỤC HÌNH ẢNH

LỜI NÓI ĐẦU

Trong thời buổi hiện đại ngày nay, công nghệ thông tin cũng như những ứng dụng của nó không ngừng phát triển, lượng thông tin và cơ sở dữ liệu được thu thập và lưu trữ cũng tích lũy ngày một nhiều lên. Con người cũng vì thế mà cần có thông tin với tốc độ nhanh nhất để đưa ra quyết định dựa trên lượng dữ liệu khổng lồ đã có. Các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế. Vì thế, một khuynh hướng kỹ thuật mới là kỹ thuật phát hiện tri thức và khai phá dữ liệu nhanh chóng được phát triển.

Khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới. Ở Việt Nam, kỹ thuật này đang được nghiên cứu và dần đưa vào ứng dụng. Khai phá dữ liệu là một bước trong quy trình phát hiện tri thức. Hiện nay, mọi người không ngừng tìm tòi các kỹ thuật để thực hiện khai phá dữ liệu một cách nhanh nhất và có được kết quả tốt nhất.

Trong bài tập lớn này, chúng em tìm hiểu và trình bày về một kỹ thuật trong khai phá dữ liệu để phân lớp dữ liệu cũng như tổng quan về khai phá dữ liệu, với đề tài "Ứng dụng thuật toán K-MEAN để phân cụm khách hàng của siêu thị". Trong quá trình làm bài tập lớn này, chúng em xin gửi lời cảm ơn đến thầy Vũ Văn Định. Thầy đã rất tận tình hướng dẫn chi tiết cho chúng em, những kiến thức thầy cung cấprất hữu ích. Chúng em rất mong nhận được những góp ý từ thầy.

Chúng em xin chân thành cảm ơn

CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1. Khái niệm khai phá dữ liệu là gì?

Data mining – khai phá dữ liệu, là một tập hợp, một hệ thống các phương pháp tính toán, thuật toán được áp dụng cho các cơ sở dữ liệu lớn và phức tạp mục đích loại bỏ các chi tiết ngẫu nhiên, chi tiết ngoại lệ, khám phá các mẫu, mô hình, quy luật tiềm ẩn, các thông tin có giá trị trong bộ dữ liệu. Data mining là thành quả công nghệ tiên tiến ngày nay, là quá trình khám phá các kiến thức vô giá bằng cách phân tích khối lượng lớn dữ liệu đồng thời lưu trữ chúng ở nhiều cơ sở dữ liệu khác nhau".

1.2. Quy trình khai phá dữ liệu

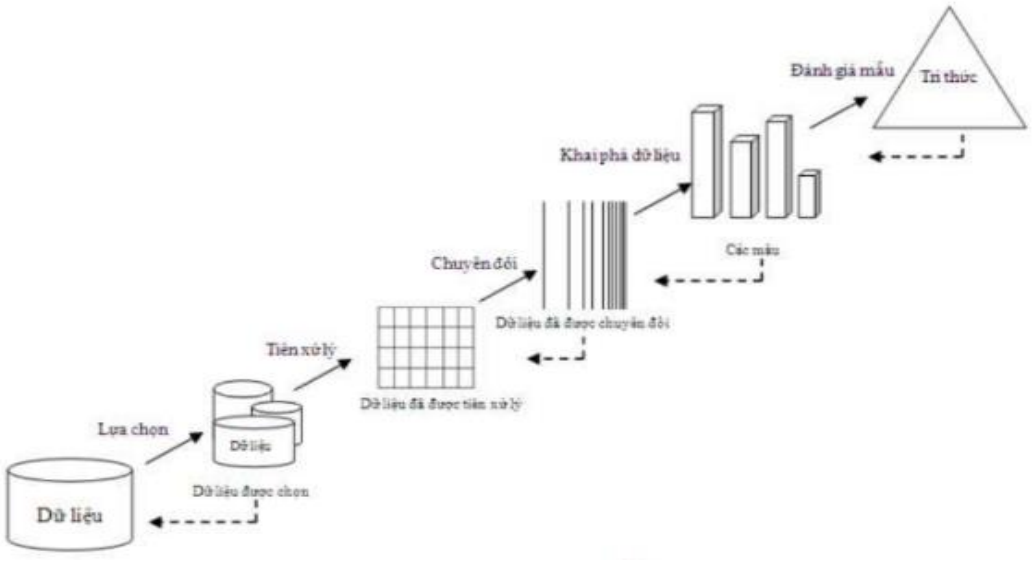
Data mining là quá trình phân tích dữ liệu để khám phá mẫu, thông tin tiềm ẩn và kiến thức hữu ích từ các tập dữ liệu lớn. Quá trình này thường được chia thành các giai đoạn cơ bản sau:

- Thu thập dữ liệu (Data Collection):
    - Thu thập dữ liệu từ nhiều nguồn khác nhau như cơ sở dữ liệu, tệp tin, trang web, hoặc thiết bị cảm biến.
    - Dữ liệu có thể là cấu trúc (như cơ sở dữ liệu quan hệ) hoặc không cấu trúc (như dữ liệu văn bản, hình ảnh, âm thanh).
- Tiền xử lý dữ liệu (Data Preprocessing):
    - Loại bỏ dữ liệu nhiễu và thiếu sót.
    - Chuẩn hóa dữ liệu để đảm bảo các đặc trưng có cùng đơn vị
    - Chuyển đổi dữ liệu không cấu trúc thành dữ liệu cấu trúc.
- Mô hình hóa và khai phá dữ liệu (Data Modeling and Exploration):
    - Chọn mô hình phù hợp để khám phá dữ liệu, như cây quyết định, mạng nơ-ron, hoặc thuật toán phân cụm.
    - Áp dụng các thuật toán data mining để khám phá mẫu, quy luật, hay cấu trúc tiềm ẩn trong dữ liệu.

- Đánh giá và tinh chỉnh các mô hình để cải thiện hiệu suất và khả năng diễn giải.

- Đánh giá kết quả (Evaluation):
  - Đánh giá hiệu suất của mô hình sử dụng các phương pháp như cross-validation, holdout sets, hoặc kiểm định thống kê.
  - So sánh các mô hình và chọn ra mô hình tốt nhất dựa trên các tiêu chí như độ chính xác, độ phủ, hay sự diễn giải.
- Triển khai (Deployment):
  - Triển khai mô hình đã huấn luyện vào môi trường thực tế để áp dụng vào các tác vụ phân loại, dự đoán, hay khuyến nghị.
  - Đảm bảo tính ổn định và hiệu suất của mô hình trong môi trường sản xuất.

Mỗi giai đoạn trong quá trình data mining đều quan trọng và đóng vai trò không thể thiếu để đạt được kết quả tốt.



1.3. Một số kỹ thuật khai phá dữ liệu

1.3.1. Kỹ thuật khai phá kết hợp

Trong khai phá dữ liệu, mục đích của luật kết hợp là tìm ra các mối quan hệ giữa các đối tượng trong khói lượng lớn dữ liệu. Để khai phá dữ liệu kết hợp có rất nhiều thuật toán nhưng phổ biến nhất là thuật toán Apriori. Đây là thuật toán khai phá tập hợp phổ biến trong dữ liệu giao dịch để phát hiện các luật kết hợp dạng khẳng định nhị phân và được sử dụng để xác định, tìm ra các luật kết hợp trong dữ liệu giao dịch. Ngoài ra, còn có các thuật toán FP-growth, thuật toán Partition,…

1.3.2. Kỹ thuật phân lớp

Trong kỹ thuật phân lớp gồm có các thuật toán:

- Phân lớp bằng cây quyết định (giải thuật ID3, J48): phân lớp dữ liệu dựa trên việc lập nên cây quyết định, nhìn vào cây quyết định có thể ra quyết định dữ liệu thuộc phân lớp nào.
- Phân lớp dựa trên xác suất (Naïve Bayesian): dựa trên việc giả định các thuộc tính độc lập mạnh với nhau qua việc sử dụng định lý Bayes.
- Phân lớp dựa trên khoảng cách (giải thuật K – láng giềng): làm như láng giềng làm, dữ liệu sẽ được phân vào lớp của k đối tượng gần với dữ liệu đó nhất.
- Phân lớp bằng SVM: phân lớp dữ liệu dựa trên việc tìm ra một siêu phẳng“tốt nhất” để tách các lớp dữ liệu trên không gian nhiều chiều hơn.

1.3.3. Kỹ thuật phân cụm

Phân cụm dữ liệu là cách phân bố các đối tượng dữ liệu vào các nhóm / cụm sao cho các đối tượng trong một cụm thì giống nhau hơn các phần tử khác cụm, gồm có một số phương pháp phân cụm cơ bản như:

- Phân cụm bằng phương pháp K-mean: tìm ra tâm của các cụm mà khoảng cách của tâm đó đến các đối tượng, dữ liệu khác là ngắn.

- Phân cụm trên đồ thị.

Ngoài ra, khai phá dữ liệu có rất nhiều kỹ thuật, nhưng đây là những kỹ thuật cơ bản và đơn giản trong khai phá dữ liệu mà chúng em tìm hiểu được.

1.4. Ứng dụng của khai phá dữ liệu

Khai phá dữ liệu có nhiều ứng dụng quan trong trong nhiều lĩnh vực khác nhau như:

- Kinh doanh và tiếp thị:
    - Dự đoán xu hướng thị trường và đánh giá hiệu suất sản phẩm.
    - Xác định nhóm đối tượng khách hàng tiềm năng dựa trên hành vi mua hàng trước đây.
    - Tối ưu hóa chiến lược giá cả và quảng cáo dựa trên dữ liệu tiêu dùng.
- Trong y tế:
    - Phát hiện các mẫu trong dữ liệu y tế để dự đoán bệnh lý và chuẩn đoán bệnh.
    - Dự đoán nguy cơ bệnh tật và xác định các nhóm nguy cơ cao để thực hiện can thiệp sớm.
    - Quản lý tài nguyên y tế và tối ưu hóa quy trình chăm sóc sức khỏe.
- Trong tài chính:
    - Dự đoán biến động thị trường tài chính và phân tích rủi ro đầu tư.
    - Phát hiện gian lận và hoạt động tài chính bất thường.
    - Đánh giá khả năng thanh toán của khách hàng và quản lý rủi ro tín dụng.
- Chăm sóc khách hàng:
    - Tạo ra hệ thống khuyến mãi và chương trình thưởng dựa trên hành vi mua hàng của khách hàng.
    - Phân loại và đánh giá mức độ hài lòng của khách hàng để cải thiện dịch vụ và tăng cường trải nghiệm khách hàng.

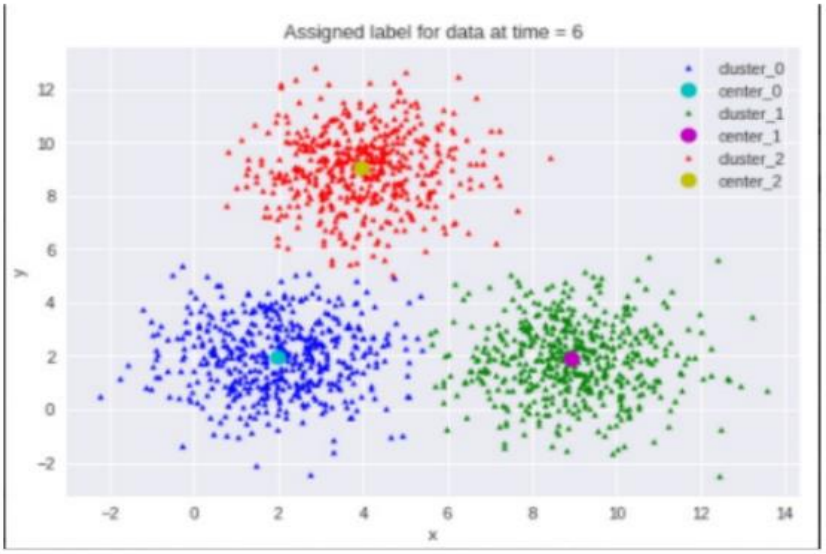
- Khoa học và nghiên cứu
    - Tạo ra hệ thống khuyến mãi và chương trình thưởng dựa trên hành vi mua hàng của khách hàng.
    - Phân loại và đánh giá mức độ hài lòng của khách hàng để cải thiện dịch vụ và tăng cường trải nghiệm khách hàng.
- An ninh mạng:
    - Tạo ra hệ thống khuyến mãi và chương trình thưởng dựa trên hành vi mua hàng của khách hàng.
    - Phân loại và đánh giá mức độ hài lòng của khách hàng để cải thiện dịch vụ và tăng cường trải nghiệm khách hàng.

Những ứng dụng này chỉ là một phần nhỏ của những gì có thể làm được với khai phá dữ liệu và mỗi lĩnh vực đều có những ứng dụng cụ thể riêng phù hợp với yêu cầu và mục tiêu cụ thể.

CHƯƠNG 2: THUẬT TOÁN K-MEAN

2.1. Giới thiệu thuật toán K-Mean

Thuật toán K-means (phân cụm K-means) nó thuộc lớp phương pháp học không giám sát. Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuâth toán phân cụm đều sinh ra cụm. Tuy nhiên không có tiêu chí nào là được xem là tốt nhất để đánh giá hiệu quả của phân cụm, điều này phụ thuộc vào mục đích của phân cụm như data reduction, "natural clusters", "useful" clusters, outlier detection. K-means là thuật toán rất quan trọng và được sử dụng phổ biến trong phân cụm. Tư tưởng chính của thuật toán K-means là tìm cách phân nhóm các đối tượng đã cho vào K cụm (K là số các cụm được định trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến nhóm (centroid) là nhỏ nhất.



Hình 2.1. Một ví dụ về sử dụng thuật toán K-mean

2.2. Các bước thực hiện thuật toán K-mean

Thuật toán k-means sử dụng phương pháp tạo và cập nhật trung tâm để phân nhóm các điểm dữ liệu cho trước vào các nhóm khác nhau. Đầu tiên chúng sẽ tạo ra các điểm trung tâm ngẫu nhiên. Sau đó gán mỗi điểm trong tập dữ liệu vào trung tâm gần nó nhất. Sau đó chúng sẽ cập nhật lại trung tâm và tiếp tục lặp lại các bước đã kể trên. Điều kiện dừng của thuật toán: Khi các trung tâm không thay đổi trong 2

vòng lặp kế tiếp nhau. Tuy nhiên, việc đạt được 1 kết quả hoàn hảo là rất khó và rất tốn thời gian, vậy nên thường người ta sẽ cho dừng thuật toán khi đạt được 1 kết quả gần đúng và chấp nhận được.

Các bước thưch hiện thuật toán:

- Bước 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.

$$C^{(0)} = \{m_1^{(0)}, m_2^{(0)}, \dots, m_k^{(0)}\}$$

- Bước 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. Tập hợp các điểm được gán vào cùng một trung tâm sẽ tạo thành cụm.

$$\mathbb{S}_i^{(t)} = \left\{ x_p : \| x_p - m_i^{(t)} \|^2 \leq \| x_p - m_j^{(t)} \|^2 \right\}, \ \forall j, \ 1 \leq j \leq k$$

- Bước 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.

$$m_i^{(t+1)} = \frac{1}{\left|\mathbb{S}_i^{(t)}\right|} \sum_{x \in \mathbb{S}_i^{(t)}} x_j$$

- Bước 4: Bước 2 và bước 3 sẽ được lặp đi lặp lại cho tới khi vị trí của tâm cụmkhông thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

Trên thực tế, sẽ có một vài lưu ý cần phải giải quyết khi áp dụng thuật toán k-mean.

2.3. Đặc điểm của thuật toán K-means

Bài toán tối ưu hóa: Cực trị cục bộ

Mỗi cụm được đặc trưng hóa bởi trung tâm của cụm (i.e. đối tượng trung bình (mean)).

- Không thể xác định được đối tượng trung bình
- Số cụm k nên là bao nhiêu?

Độ phức tạp O(nkt). Trong đó:

- n là số đối tượng, k là số cụm, t là số lần lặp
- k << n, t << n

Ảnh hưởng bởi nhiễu (các phần tử kì dị/biên)

Không phù hợp cho việc khai phá ra các cụm có dạng không lồi (nonconvex) hay các cụm có kích thước rất khác nhau

- Kết quả gom cụm có dạng siêu cầu (hyperspherial).
- Kích thước các cụm kết quả thường đồng đều (relatively uniform sizes).

2.4. Nhận xét, đánh giá thuật toán

Một số ưu điểm của thuật toán phân cụm K-means:

- Dễ hiểu và dễ thực hiện.
- Nếu chúng ta có số lượng biến lớn thì K-mean sẽ nhanh hơn so với phân cụm phâncấp.
- 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.

Một số nhược điểm của thuật toán:

- 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 mìnhbằng 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

CHƯƠNG 3: CÀI ĐẶT THUẬT TOÁN

3.1. Dữ liệu đầu vào

Dữ liệu đầu vào ta có dữ liệu khách hàng của một siêu thị được lưu dưới dạng file csv bao gồm 5 cột và 200 hàng

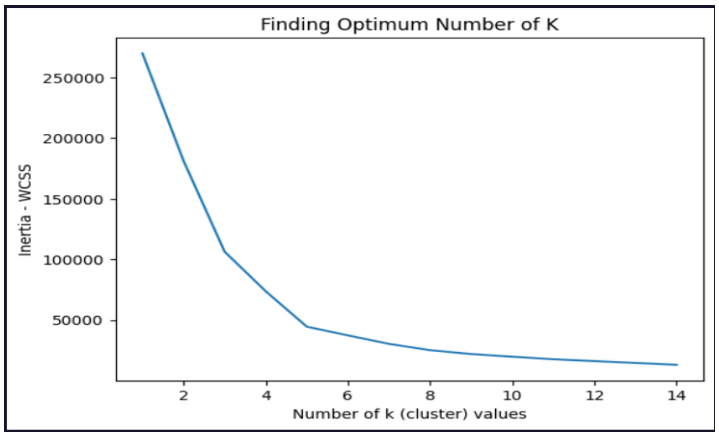
Dưới đây là một số dữ liệu có trong file dữ liệu được sử dụng

CustomerID	Gender	Age	Annual_Income_(k$)	Spending_Score
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35
18	Male	20	21	66
19	Male	52	23	29

Hình 3.1. Một số dữ liệu được sử dụng

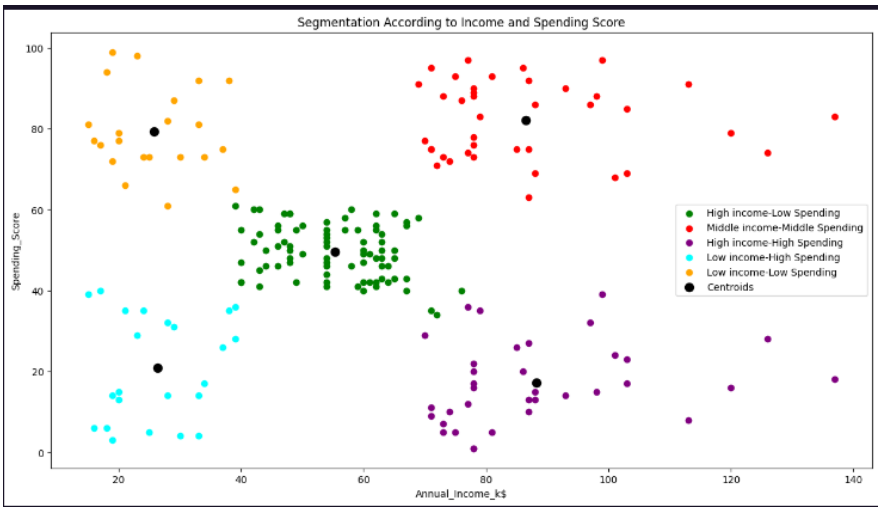
3.2. Demo chương trình

Tìm số k tối ưu



Hình 3.2. Tìm kiếm số k tối ưu

Biểu đồ phân cụm



Hình 3.3. Biểu đồ phân cụm

3.3. Tiến hành phân cụm

Import các thư viện cần thiết sau đó tiến hành đọc file dữ liệu và lưu vào dataframe

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans

df = pd.read_csv('Customers.csv')
print(df)
```

Tiến hành tìm số k tối ưu nhất và sử dụng thuật toán kmeans để phân loại dữ liệu

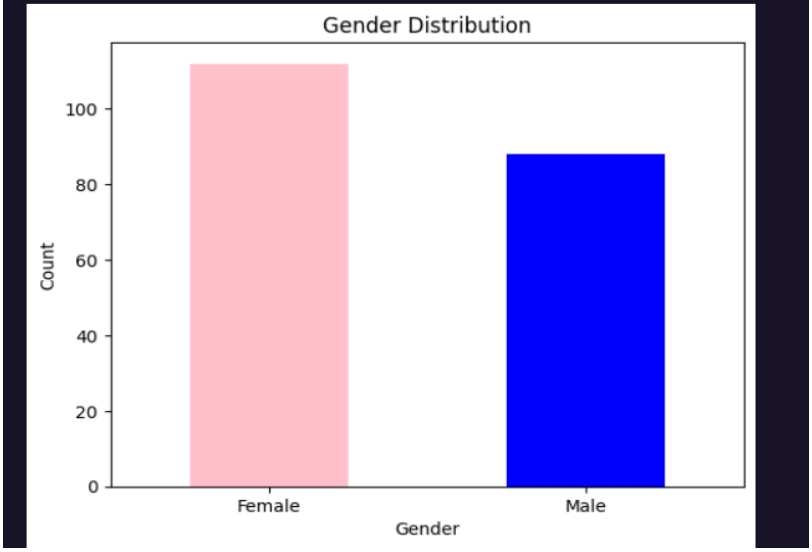
```
x_data1 = df[['Annual_Income_(k$)', 'Spending_Score']].values
lst = []
for k in range(1, 15):
    kmeans = KMeans(n_clusters=k, random_state=0)
    kmeans.fit(x_data1)
    lst.append(kmeans.inertia_)
plt.plot(range(1, 15), lst)
plt.xlabel("Number of k (cluster) values")
plt.ylabel("Inertia - WCSS")
plt.title("Finding Optimum Number of K")
plt.show()
# Áp dụng KMeans với số cụm đã chọn
kmeans_x_data1 = KMeans(n_clusters=5, random_state=0)
clusters = kmeans_x_data1.fit_predict(x_data1)
df["Label1"] = clusters

# Vẽ biểu đồ scatter plot với các cụm đã được phân loại
plt.figure(figsize=(15, 8))
plt.scatter(x_data1[clusters == 0, 0], x_data1[clusters == 0, 1], color="green", label="High income-Low Spending")
plt.scatter(x_data1[clusters == 1, 0], x_data1[clusters == 1, 1], color="red", label="Middle income-Middle Spending")
plt.scatter(x_data1[clusters == 2, 0], x_data1[clusters == 2, 1], color="purple", label="High income-High Spending")
plt.scatter(x_data1[clusters == 3, 0], x_data1[clusters == 3, 1], color="cyan", label="Low income-High Spending")
plt.scatter(x_data1[clusters == 4, 0], x_data1[clusters == 4, 1], color="orange", label="Low income-Low Spending")
plt.scatter(kmeans_x_data1.cluster_centers_[:, 0], kmeans_x_data1.cluster_centers_[:, 1], color="black", label="Centroids", s=7
plt.xlabel("Annual_Income_k$")
plt.ylabel("Spending_Score")
plt.legend()
plt.title("Segmentation According to Income and Spending Score")
plt.show()
```

Biểu đồ phân loại giới tính (Gender)

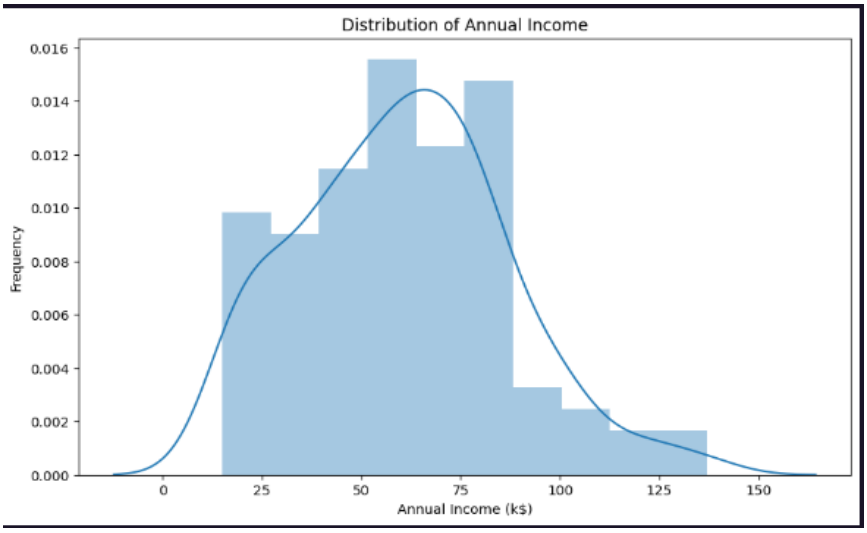
```
# Tạo biểu đồ phân loại g
gender_counts = df['Gender'].value_counts()
gender_counts.plot(kind='bar', color=
['pink', 'blue'])
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.show()
```

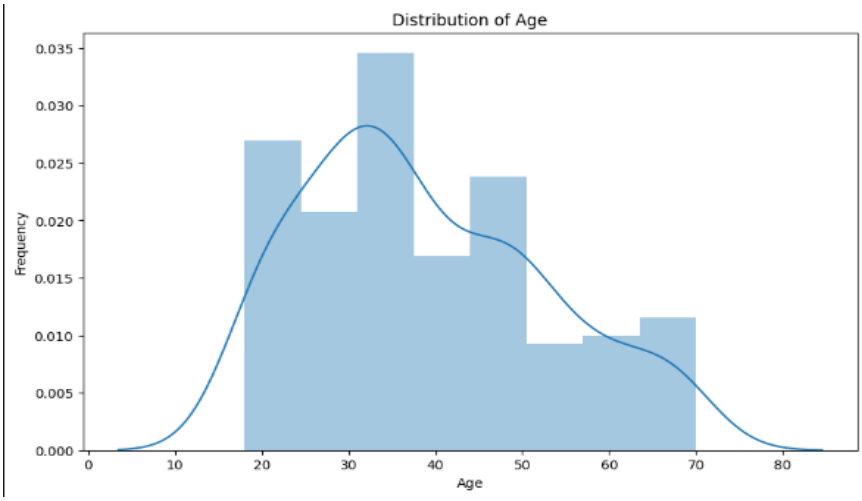
✓ 0.1s Python

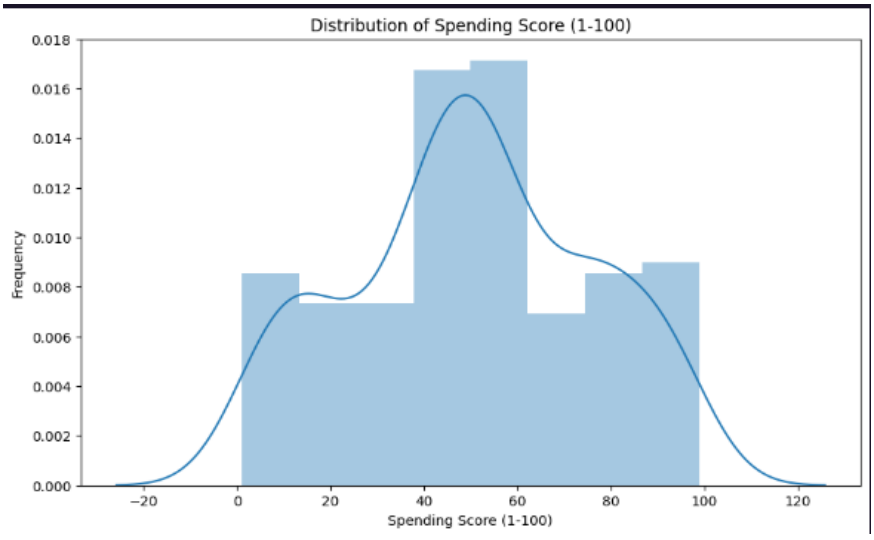


Đồng thời là biểu đồ thể hiện thu nhập, độ tuổi, điểm chi tiêu của khách hàng.

Kết quả:







KẾT LUẬN

Trong quá trình thực hiện bài báo cáo của học phần khai phá dữ liệu, chúng em đã được biết thêm về các chương trình ứng dụng, nắm rõ các phần về khai phá dữ liệu. Từ đó chúng em cố gắng áp dụng các kiến thức đã học được vào làm báo cáo để hoàn tiện sản phẩm của mình. Kết quả đã đạt được khi kết thúc học phần và hoàn thiện bài báo cáo:

- Tìm hiểu về học máy và các bài toán trong học máy
- Tìm hiểu về nhiều thuật toán như Navie Bayes, Decision Tree, Random ForestClassification, K-Means,..
- Xây dựng được cơ bản mô hình phân cụm và dự đoán.

Vì thời gian triển khai có hạn, việc tìm hiểu công nghệ mới còn gặp nhiều khó khăn do không có tài liệu cụ thể nên không tránh khỏi sai sót.

TÀI LIỆU THAM KHẢO

[1] https://vi.wikipedia.org/wiki/Ph%C3%A2n_c%E1%BB%A5m_k-means

[2] https://tigosoftware.com/vi/thuat-toan-k-means-voi-bai-toan-phan-cum-du-lieu