

TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

KHOA ĐIỆN TỬ

Bộ môn: Công nghệ Thông tin.

BÀI TẬP LỚN

MÔN HỌC

KHOA HỌC DỮ LIỆU

Sinh viên: LẠI CHÍ BẢO

Lớp: K57KMT

Giáo viên hướng dẫn: NGUYỄN VĂN HUY

Link GitHub: https://github.com/baolaichi/BTL_KhoaHocDuLieu_LaiChiBao.git



Thái Nguyên – 2025

TRƯỜNG ĐHKTCN
KHOA ĐIỆN TỬ

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÀI TẬP LỚN

MÔN HỌC: KHOA HỌC DỮ LIỆU
BỘ MÔN: CÔNG NGHỆ THÔNG TIN

Sinh viên: LẠI CHÍ BẢO

Lớp: 57KMT

Ngành: KỸ THUẬT MÁY TÍNH

Giáo viên hướng dẫn: NGUYỄN VĂN HUY

Ngày giao đề..... Ngày hoàn thành.....

Tên đề tài: Phân loại khách hàng tiềm năng

Yêu cầu:

Đầu bài:

Ứng dụng web phân loại khách hàng dựa trên dữ liệu giao dịch.

Đầu vào:

[Mall Customer Segmentation Data](#)

Đầu ra:

Phân loại khách hàng và biểu đồ phân nhóm khách hàng.

GIÁO VIÊN HƯỚNG DẪN

(Ký và ghi rõ họ tên)

This image shows a full page of white paper with horizontal dotted lines. The lines are evenly spaced and run across the width of the page, providing a guide for handwriting or typing. There are no margins, text, or other markings on the page.

GIÁO VIÊN HƯỚNG DẪN

3

Mục Lục

MỞ ĐẦU	5
CHƯƠNG 1: GIỚI THIỆU ĐẦU BÀI.....	6
1.1. Đặt vấn đề	6
1.2. Yêu cầu và chức năng chính	6
1.3. Thách thức	6
1.4. Kiến thức áp dụng	7
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT	8
2.1. Xử lý dữ liệu với Pandas	8
2.2. Trực quan hóa dữ liệu	9
2.3. Phân cụm khách hàng bằng KMeans	11
2.4. Dự đoán mức chi tiêu với mô hình học máy	11
CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH	14
3.1. Sơ đồ khối hệ thống	14
3.4. Chương trình	17
CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT LUẬN	18
4.1. Thực nghiệm	18
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	21
5.3. Hướng phát triển	22
TÀI LIỆU THAM KHẢO	23

MỞ ĐẦU

Trong bối cảnh cạnh tranh ngày càng gay gắt của thị trường hiện đại, việc hiểu rõ nhu cầu và hành vi tiêu dùng của khách hàng là yếu tố then chốt giúp các doanh nghiệp xây dựng chiến lược kinh doanh hiệu quả. Phân tích dữ liệu khách hàng không chỉ giúp doanh nghiệp phân loại khách hàng mà còn hỗ trợ dự đoán xu hướng tiêu dùng, từ đó tập trung nguồn lực vào các nhóm khách hàng tiềm năng nhất.

Đề tài “**Phân loại khách hàng tiềm năng dựa trên dữ liệu giao dịch**” là một ứng dụng thực tiễn của khoa học dữ liệu (Data Science) trong lĩnh vực marketing và bán lẻ. Thông qua việc xử lý dữ liệu từ tập dữ liệu *Mall Customer Segmentation*, nhóm chúng em xây dựng một ứng dụng web có khả năng:

- Xử lý và trực quan hóa dữ liệu khách hàng
- Phân cụm khách hàng theo các đặc trưng như tuổi, giới tính và thu nhập
- Dự đoán mức tiêu dùng của khách hàng mới bằng mô hình học máy
- Hiển thị kết quả phân tích và dự đoán trên giao diện web thân thiện

Ứng dụng được xây dựng với các công nghệ hiện đại như **Python, Flask, Pandas, scikit-learn, Plotly**, kết hợp giao diện Bootstrap để đảm bảo cả tính năng lẫn thẩm mỹ.

Báo cáo này sẽ trình bày chi tiết quy trình xử lý dữ liệu, xây dựng mô hình học máy, thiết kế giao diện và đánh giá hiệu quả mô hình trong việc hỗ trợ phân loại và dự đoán khách hàng tiềm năng.

Chúng em xin chân thành cảm ơn thầy đã hướng dẫn và tạo điều kiện để nhóm thực hiện đề tài này. Mặc dù đã cố gắng hoàn thiện trong khả năng cho phép, nhưng không thể tránh khỏi thiếu sót, kính mong thầy góp ý để nhóm có thể học hỏi và cải thiện thêm.

CHƯƠNG 1: GIỚI THIỆU ĐẦU BÀI

1.1. Đặt vấn đề

Trong thời đại chuyển đổi số, dữ liệu khách hàng ngày càng trở thành tài sản quan trọng đối với các doanh nghiệp. Việc phân tích và khai thác dữ liệu một cách hiệu quả không chỉ giúp doanh nghiệp thấu hiểu nhu cầu khách hàng mà còn góp phần nâng cao hiệu quả kinh doanh thông qua các chiến lược tiếp thị, chăm sóc và giữ chân khách hàng phù hợp.

Trong môn học *Data Science*, bài tập lớn được giao với đề tài:

"Phân loại khách hàng tiềm năng dựa trên dữ liệu giao dịch", sử dụng tập dữ liệu **Mall Customer Segmentation Data**. Mục tiêu là xây dựng một ứng dụng web có khả năng xử lý dữ liệu khách hàng, thực hiện phân cụm (clustering), dự đoán mức tiêu dùng, và trực quan hóa kết quả trên giao diện người dùng.

1.2. Yêu cầu và chức năng chính

Ứng dụng cần đáp ứng các chức năng chính sau:

- Đọc và xử lý dữ liệu khách hàng từ tập dữ liệu có sẵn bằng thư viện **Pandas**
- Xử lý dữ liệu khuyết thiếu, chuẩn hóa các đặc trưng
- **Phân cụm khách hàng** sử dụng thuật toán **KMeans** để nhóm khách hàng theo các đặc trưng như độ tuổi, thu nhập và điểm tiêu dùng
- **Dự đoán mức chi tiêu (Spending Score)** của khách hàng mới thông qua mô hình **Random Forest Regression**
- Trực quan hóa dữ liệu và kết quả phân tích bằng đồ thị **Histogram**, **Scatter Plot**, và biểu đồ **3D** sử dụng **Matplotlib**, **Seaborn** và **Plotly**
- Cho phép người dùng **nhập thông tin khách hàng mới** để dự đoán mức tiêu dùng
- Giao diện đẹp, thân thiện, sử dụng **Flask** và **Bootstrap**
- Cho phép **xuất kết quả ra file CSV**

1.3. Thách thức

Một số thách thức khi thực hiện đề tài bao gồm:

- Việc lựa chọn và xử lý các đặc trưng phù hợp cho mô hình học máy (feature engineering)
- Xử lý dữ liệu không đầy đủ hoặc bị thiếu cột (thường gặp trong tập dữ liệu thực tế)
- Đánh giá và lựa chọn mô hình dự đoán phù hợp giữa **Linear Regression** và **Random Forest Regression**
- Thiết kế giao diện web vừa trực quan, vừa dễ sử dụng để hiển thị kết quả phân tích

- Hiện thị biểu đồ phân cụm, phân phối một cách sinh động, có tương tác

1.4. Kiến thức áp dụng

Để thực hiện đề tài này, nhóm đã vận dụng tổng hợp các kiến thức đã học trong môn *Data Science* và các công cụ liên quan, bao gồm:

- **Tiền xử lý và phân tích dữ liệu:** sử dụng Pandas, NumPy
- **Học máy (Machine Learning):** thuật toán KMeans, Linear Regression, Random Forest Regression (scikit-learn)
- **Trực quan hóa dữ liệu:** sử dụng Matplotlib, Seaborn, Plotly
- **Phát triển web:** sử dụng Flask (Python), HTML/CSS và Bootstrap
- **Triển khai mô hình ML vào ứng dụng thực tế**

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

Trong quá trình thực hiện đề tài, đã vận dụng nhiều kiến thức liên quan đến xử lý dữ liệu, học máy và phát triển ứng dụng web. Dưới đây là tổng hợp những cơ sở lý thuyết và công cụ đã sử dụng:

2.1. Xử lý dữ liệu với Pandas

2.1.1. Khái quát về Pandas

Pandas là một thư viện mạnh trong Python, được sử dụng phổ biến trong lĩnh vực khoa học dữ liệu. Nó hỗ trợ thao tác và xử lý dữ liệu dạng bảng (DataFrame) một cách hiệu quả.

2.1.2. Các thao tác cơ bản sử dụng trong chương trình

Trong đề tài "**Phân loại khách hàng tiềm năng dựa trên dữ liệu giao dịch**", thư viện **Pandas** được sử dụng chủ yếu để xử lý và tiền xử lý dữ liệu. Cụ thể, các bước như sau:

a) Đọc dữ liệu từ file CSV

Dữ liệu khách hàng được lưu trữ trong file `Mall_Customers.csv`, bao gồm các cột như `CustomerID`, `Genre`, `Age`, `Annual Income (k$)` và `Spending Score (1-100)`.

Trong chương trình, dữ liệu được đọc bằng dòng lệnh:

```
df = pd.read_csv('Mall_Customers.csv')
```

b) Kiểm tra dữ liệu bị thiếu

Trước khi đưa dữ liệu vào huấn luyện mô hình, cần đảm bảo không có giá trị thiếu (NaN). Chương trình kiểm tra bằng câu lệnh:

```
df.isnull().sum()
```

Kết quả được hiển thị ra bảng, giúp phát hiện nếu có bất kỳ cột nào chứa giá trị bị thiếu. Với dữ liệu gốc của đề tài, không có giá trị thiếu nên không cần xử lý thêm bước này. Tuy nhiên, câu lệnh trên được giữ để đảm bảo tính tổng quát và sẵn sàng nếu dữ liệu mới có lỗi.

c) Thay thế giá trị thiếu hoặc loại bỏ dòng/đặc trưng không cần thiết

Trong trường hợp có giá trị thiếu, có thể dùng `fillna()` để thay thế hoặc `dropna()` để loại bỏ. Ví dụ:

```
df.dropna(inplace=True) # Nếu có dòng bị thiếu dữ liệu
```

Ngoài ra, các cột không ảnh hưởng đến mô hình như `CustomerID` cũng có thể được loại bỏ khỏi tập huấn luyện để giảm nhiễu:

```
df = df.drop(['CustomerID'], axis=1)
```

d) Biến đổi dữ liệu dạng chuỗi thành số

Trong dữ liệu gốc, cột `Gender` có kiểu dữ liệu là chuỗi (`Male`, `Female`). Để mô hình học máy xử lý được, cần biến đổi sang dạng số. Chương trình sử dụng `.map()` như sau:

```
df['Gender'] = df['Gender'].map({'Male': 0, 'Female': 1})
```

Phương pháp này chuyển giá trị `Male` thành 0 và `Female` thành 1. Việc mã hóa nhị phân này là phù hợp vì `Gender` chỉ có hai giá trị phân biệt, không cần dùng `OneHotEncoder`.

2.2. Trục quan hóa dữ liệu

2.2.1. Mục đích

Việc trục quan hóa giúp hiểu rõ mối quan hệ giữa các đặc trưng và xu hướng trong tập dữ liệu. Đây là bước quan trọng để hỗ trợ chọn lựa đặc trưng cho mô hình máy học.

2.2.2. Các thư viện sử dụng

- a) **Matplotlib** – hỗ trợ tạo các biểu đồ dạng cơ bản (line, bar, scatter, histogram)
- b) **Seaborn** – xây dựng biểu đồ đẹp hơn, tích hợp thống kê
- c) **Plotly** – thư viện tương tác, sử dụng để tạo biểu đồ 3D và biểu đồ có thể xoay được trong giao diện web

2.2.3. Các biểu đồ sử dụng

Trong đề tài, việc trực quan hóa dữ liệu đóng vai trò quan trọng nhằm giúp người dùng hiểu rõ hơn về đặc điểm khách hàng cũng như kết quả phân tích, phân loại. Ứng dụng đã tích hợp nhiều loại biểu đồ, mỗi loại phục vụ một mục đích cụ thể. Các biểu đồ này được xây dựng bằng thư viện Matplotlib, Seaborn, và Plotly, và được hiển thị trực tiếp trên giao diện web Flask.

- **Hình 1:** Biểu đồ phân phối độ tuổi khách hàng (Histogram)
- **Mục đích:** Giúp người dùng hiểu độ tuổi phổ biến của nhóm khách hàng hiện có trong cơ sở dữ liệu.
- **Thực hiện:** Sử dụng `plt.hist()` từ Matplotlib để vẽ biểu đồ tần suất.

```
plt.hist(df['Age'], bins=10, edgecolor='black')
plt.title('Phân phối độ tuổi khách hàng')
plt.xlabel('Tuổi')
plt.ylabel('Số lượng khách hàng')
```

Trong web: Biểu đồ này được hiển thị ở trang “Khám phá dữ liệu” (Data Exploration) như một phần của giao diện tổng quan. Nhờ đó, người quản trị có thể quan sát và đưa ra nhận định về phân bố tuổi trong tập dữ liệu.

- **Hình 2:** Biểu đồ phân cụm khách hàng theo 3 chiều (Plotly 3D Scatter)
- **Mục đích:** Trực quan hóa kết quả phân cụm khách hàng sau khi áp dụng thuật toán KMeans với 3 đặc trưng: Tuổi (Age), Thu nhập hàng năm (Income) và Điểm chi tiêu (Spending).
- **Thực hiện:** Sử dụng `plotly.express.scatter_3d()` để tạo biểu đồ tương tác 3 chiều.

```
fig = px.scatter_3d(clustered_df, x='Age', y='Income', z='Spending',
color='Cluster',
```

```
title='Phân cụm khách hàng theo KMeans (3D)',
```

```
labels={'Age': 'Tuổi', 'Income': 'Thu nhập', 'Spending': 'Chi tiêu'}))
```

Trong web: Biểu đồ này được hiển thị ở phần “Kết quả phân cụm” (Clustering Results). Biểu đồ hỗ trợ kéo/zoom tương tác và màu sắc giúp dễ dàng phân biệt từng cụm khách hàng. Người dùng có thể dùng biểu đồ này để đánh giá sự phân biệt giữa các nhóm khách hàng tiềm năng.

- **Hình 3:** Biểu đồ thu nhập vs. chi tiêu (Scatter Plot)

- **Mục đích:** Cho thấy mối quan hệ giữa thu nhập hàng năm và điểm chi tiêu – hai yếu tố quan trọng dùng để đánh giá mức độ tiềm năng của khách hàng.
- **Thực hiện:** Dùng `plt.scatter()` hoặc `seaborn.scatterplot()` để hiển thị.

```
sns.scatterplot(data=df, x='Income', y='Spending', hue='Gender')
plt.title('Quan hệ giữa thu nhập và điểm chi tiêu')
plt.xlabel('Thu nhập (k$)')
plt.ylabel('Điểm chi tiêu (1-100)')
```

Trong web: Biểu đồ này cũng được tích hợp trong phần “Phân tích dữ liệu” để người dùng nắm rõ tương quan giữa các biến đầu vào chính. Việc phân tích biểu đồ này có thể giúp phát hiện các nhóm khách hàng có thu nhập cao nhưng chi tiêu thấp, từ đó xác định cơ hội tiếp cận lại.

2.3. Phân cụm khách hàng bằng KMeans

2.3.1. Khái niệm

KMeans là thuật toán phân cụm không giám sát (unsupervised learning), chia dữ liệu thành k nhóm dựa trên độ tương đồng giữa các điểm dữ liệu. Thuật toán tìm trung tâm cụm (centroid) và cập nhật dần đến khi hội tụ.

2.3.2. Ứng dụng trong đề tài

Thuật toán được dùng để phân nhóm khách hàng theo các đặc trưng:

- Tuổi (Age)
- Thu nhập hàng năm (Annual Income)
- Điểm tiêu dùng (Spending Score)

2.3.3. Xác định số cụm (k)

Sử dụng phương pháp Elbow để xác định số cụm tối ưu (Hình 4: Đồ thị Elbow Curve).

2.4. Dự đoán mức chi tiêu với mô hình học máy

2.4.1. Bài toán dự đoán

Mục tiêu là xây dựng mô hình dự đoán **Spending Score** (điểm chi tiêu) của khách hàng mới dựa vào các đặc trưng:

- Tuổi (Age)
- Giới tính (Gender)
- Thu nhập hàng năm (Annual Income)

2.4.2. Thuật toán sử dụng

a) Linear Regression

- Mô hình hồi quy tuyến tính, giả định mối quan hệ tuyến tính giữa đầu vào và đầu ra.

b) Random Forest Regression

- Mô hình hồi quy phi tuyến tính, sử dụng tập hợp nhiều cây quyết định.
- Ưu điểm: chống overfitting tốt, hoạt động hiệu quả với dữ liệu phức tạp.

2.4.3. Đánh giá mô hình

- Dùng **Mean Squared Error (MSE)** hoặc **R² Score** để đánh giá độ chính xác
- **Hình 5:** Biểu đồ so sánh giá trị dự đoán và thực tế của mô hình

2.5. Giao diện người dùng với Flask và Bootstrap

2.5.1. Flask

Là một framework nhẹ cho phát triển ứng dụng web bằng Python. Cho phép xây dựng API, truyền dữ liệu giữa client và server.

2.5.2. Bootstrap

Là thư viện CSS giúp xây dựng giao diện đẹp, đáp ứng tốt trên nhiều loại thiết bị (responsive design). Trong đề tài, Bootstrap được dùng để:

- Thiết kế biểu mẫu nhập dữ liệu
- Tạo bảng kết quả kẻ đẹp, dễ nhìn
- Tùy chỉnh giao diện thành dạng hiện đại, thân thiện

2.5.3. Chức năng trên giao diện

- Nhập thông tin khách hàng mới → dự đoán mức tiêu dùng
- Xem biểu đồ phân cụm khách hàng (KMeans)
- Xem bảng dữ liệu gốc và kết quả đã xử lý

- Xuất kết quả phân tích ra file CSV

CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

3.1. Sơ đồ khối hệ thống

Hệ thống gồm các module chính sau:

a) Giao diện người dùng (Flask Web UI)

- Cho phép người dùng tương tác với hệ thống qua trình duyệt.
- Hiển thị kết quả phân tích, phân cụm, và dự đoán chi tiêu.
- Cho phép nhập dữ liệu khách hàng mới để dự đoán.

b) Xử lý dữ liệu (Data Preprocessing)

- Đọc và xử lý file CSV.
- Làm sạch dữ liệu, xử lý dữ liệu khuyết thiếu.
- Biến đổi dữ liệu định danh (giới tính) thành số.

c) Phân cụm khách hàng (Customer Clustering)

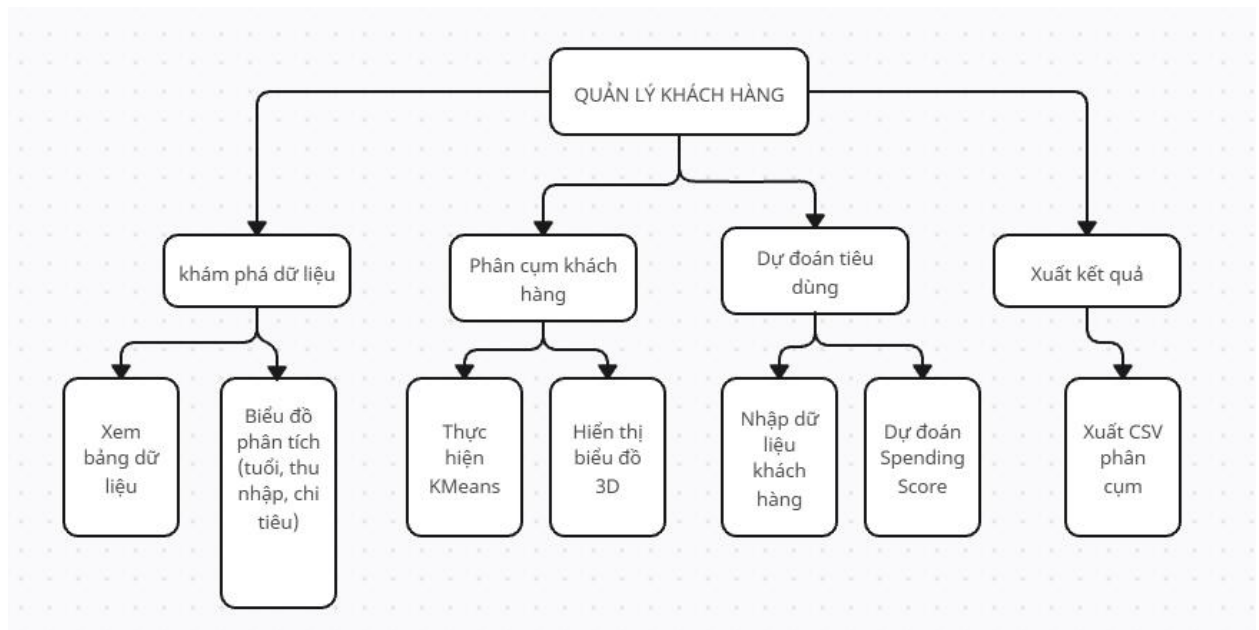
- Sử dụng KMeans để phân nhóm khách hàng theo độ tuổi, thu nhập, chi tiêu.
- Lưu kết quả phân cụm để hiển thị biểu đồ phân tích.

d) Dự đoán mức tiêu dùng (Spending Prediction)

- Cho phép người dùng nhập thông tin khách hàng mới.
- Sử dụng mô hình học máy (Random Forest) để dự đoán điểm chi tiêu (Spending Score).

e) Trực quan hóa dữ liệu (Visualization)

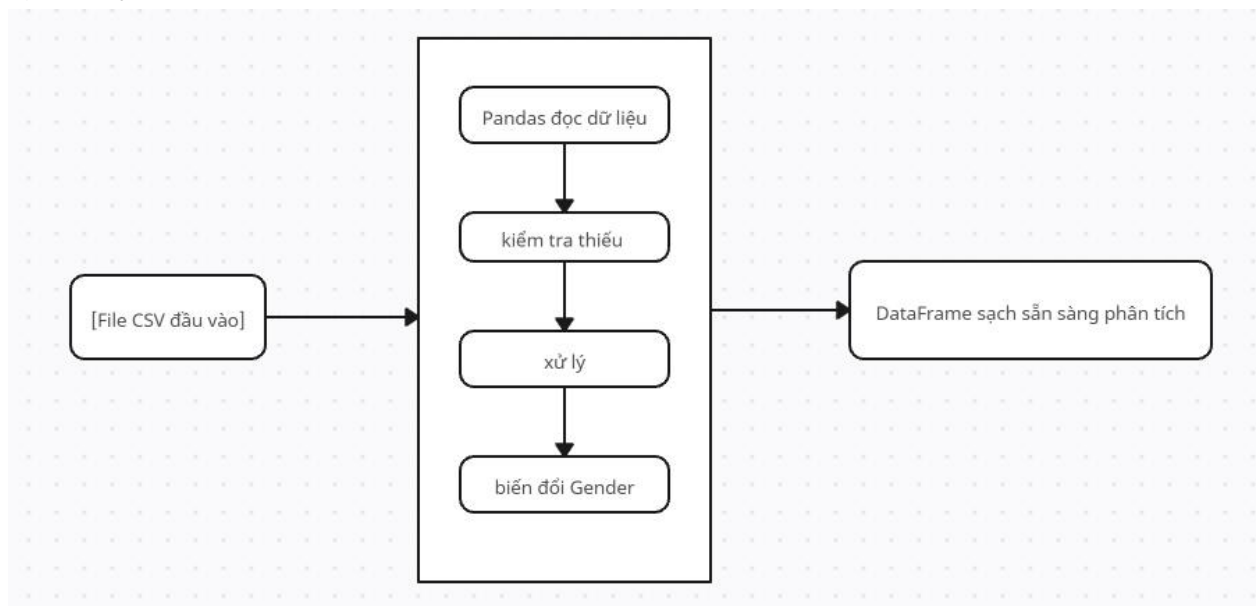
- Hiển thị biểu đồ histogram, scatter plot, và 3D scatter plot.
- Hỗ trợ người dùng trực quan hóa dữ liệu và kết quả phân tích.



Hình 1: Biểu đồ phân cấp chức năng

3.2. Sơ đồ khối các thuật toán chính

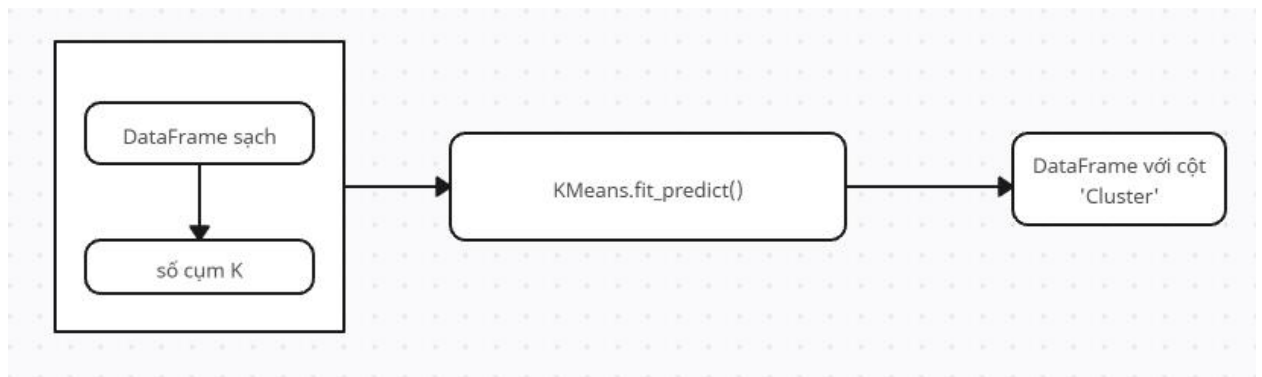
a) Xử lý dữ liệu



Hình 2: Luồng xử lý dữ liệu

- **Chức năng:** Làm sạch dữ liệu, chuyển đổi dữ liệu định danh, loại bỏ dữ liệu dư thừa.
- **Đầu vào:** File CSV chứa thông tin khách hàng.
- **Đầu ra:** DataFrame đã xử lý.

b) Phân cụm Kmeans



Hình 3: Phân cụm Kmeans

- **Chức năng:** Phân nhóm khách hàng theo đặc trưng (Tuổi, Thu nhập, Chi tiêu).
- **Đầu vào:** DataFrame đã xử lý.
- **Đầu ra:** Cột Cluster mới trong bảng dữ liệu.

c) Dự đoán mức tiêu dùng (Random Forest Regression)



Hình 4: xử lý dự đoán mức tiêu dùng

- **Chức năng:** Dự đoán điểm chi tiêu cho khách hàng mới.
- **Đầu vào:** Tuổi, giới tính, thu nhập của khách hàng mới.
- **Đầu ra:** Điểm tiêu dùng dự đoán (Spending Score).

3.3. Cấu trúc dữ liệu

Dữ liệu đầu vào là file CSV *Mall_Customers.csv* với các trường như sau:

Trường	Kiểu dữ liệu	Mô tả
CustomerID	Int	Mã định danh khách hàng
Genre	Object	Giới tính khách hàng (Male/Female)
Age	Int	Tuổi khách hàng
Annual Income (K\$)	Int	Thu nhập hàng năm
Spending Score (1-100)	Int	Điểm đánh giá tiêu dùng

Bảng 1: Bảng các trường dữ liệu đầu vào

Sau xử lý, dữ liệu bổ sung thêm cột:

Trường	Kiểu dữ liệu	Mô Tả
Gender	Int	Giới tính (Male=0, Female=1)
Cluster	Int	Cụm khách hàng sau phân nhóm

Bảng 2: Bảng trường dữ liệu bổ sung

3.4. Chương trình

Chương trình được tổ chức thành các module chính:

a) app.py

- Khởi tạo Flask app
- Định nghĩa các route: /, /explore, /cluster, /predict, /download
- Nhận dữ liệu đầu vào từ người dùng
- Gọi các hàm xử lý và trả về kết quả

b) model.py

- Chứa hàm preprocess_data() để làm sạch dữ liệu
- Hàm cluster_customers() thực hiện phân cụm KMeans
- Hàm predict_spending() huấn luyện mô hình RandomForest và trả về dự đoán chi tiêu

c) templates/

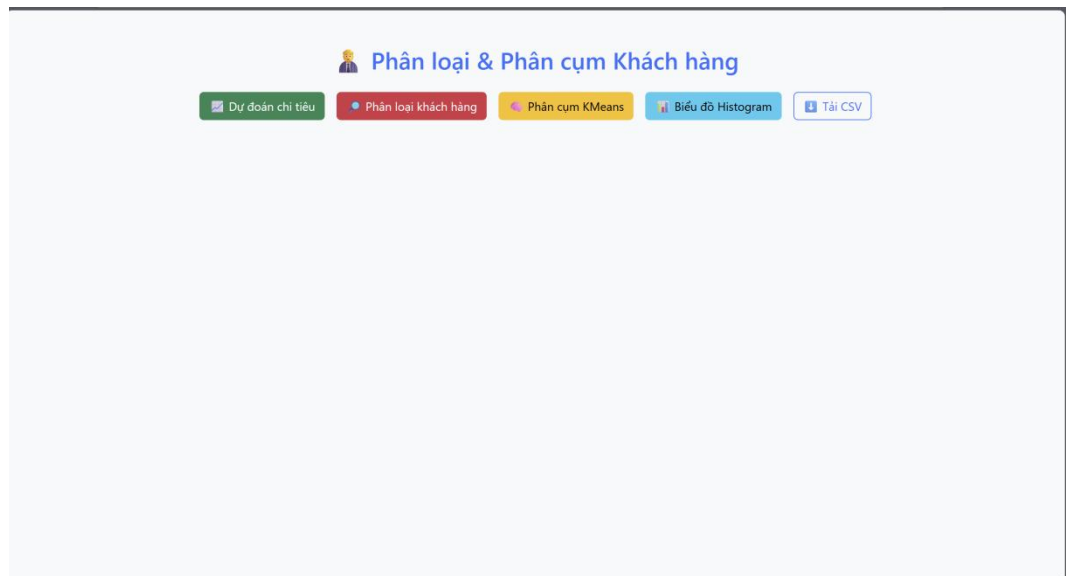
- Chứa các file HTML:
 - index.html: giao diện chính
 - explore.html: khám phá dữ liệu
 - cluster.html: hiển thị phân cụm
 - predict.html: form nhập khách hàng mới và hiển thị kết quả

d) static/

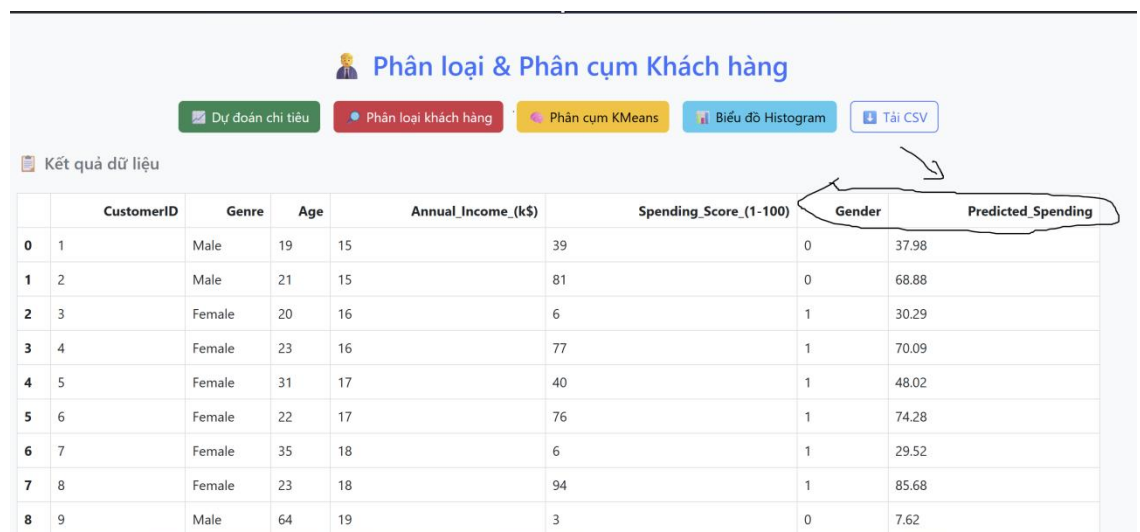
- Chứa biểu đồ đã tạo (file PNG hoặc hình ảnh từ matplotlib)

CHƯƠNG 4: THỰC NGHIỆM

4.1. Thực nghiệm

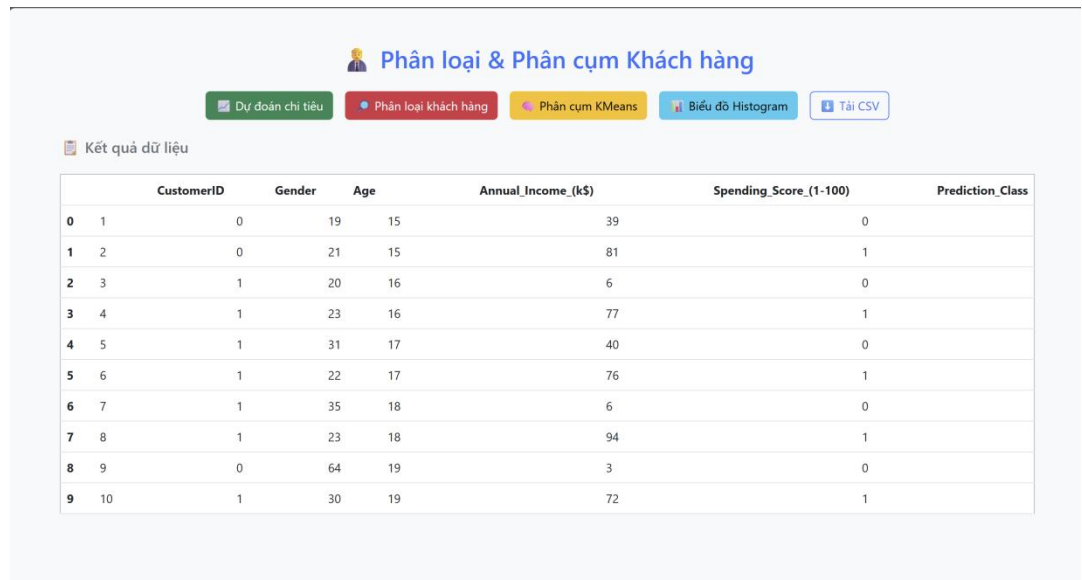


Hình 5: Giao diện Chính



	CustomerID	Genre	Age	Annual_Income_(k\$)	Spending_Score_(1-100)	Gender	Predicted_Spending
0	1	Male	19	15	39	0	37.98
1	2	Male	21	15	81	0	68.88
2	3	Female	20	16	6	1	30.29
3	4	Female	23	16	77	1	70.09
4	5	Female	31	17	40	1	48.02
5	6	Female	22	17	76	1	74.28
6	7	Female	35	18	6	1	29.52
7	8	Female	23	18	94	1	85.68
8	9	Male	64	19	3	0	7.62

Hình 6: Chức năng Dự đoán chi tiêu

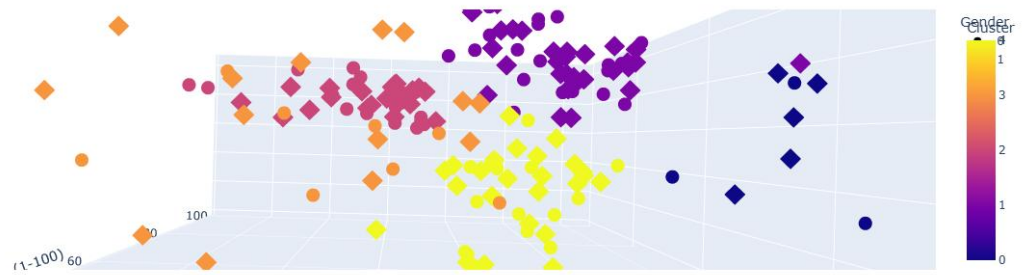


Hình 7: Chức năng phân loại khách hàng

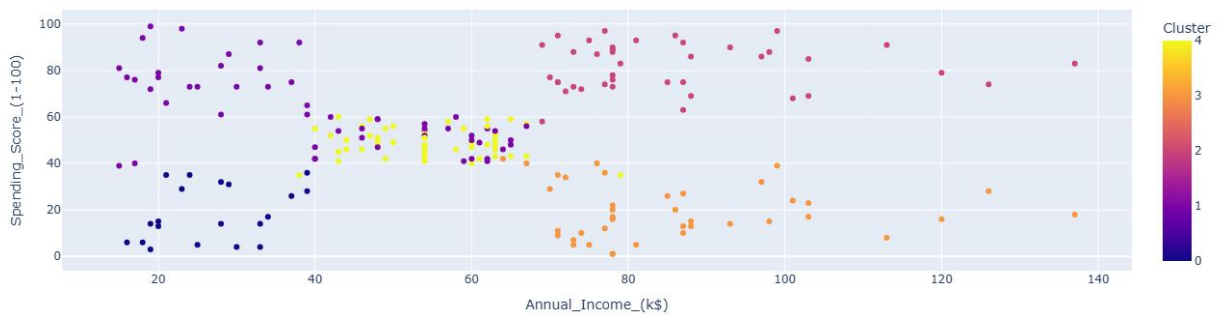
Phân cụm và phân loại dựa trên các tiêu chí sau:

1. **Độ tuổi (Age)**
 - Phân loại nhóm khách hàng trẻ (dưới 30), trung niên (30–50), và lớn tuổi (>50).
2. **Thu nhập hàng năm (Annual Income (k\$))**
 - Phân nhóm khách hàng có thu nhập thấp, trung bình, cao.
3. **Điểm chi tiêu (Spending Score)**
 - Điểm từ 1 đến 100, cho biết mức độ chi tiêu và tương tác với doanh nghiệp.
 - Dùng để nhận biết **khách hàng tiềm năng**: khách có thu nhập cao + điểm chi tiêu cao → tiềm năng.
4. **Giới tính (Gender)**
 - Được mã hóa (Male=0, Female=1) để dùng trong mô hình học máy.
5. **Phân cụm bằng KMeans**
 - Thuật toán KMeans chia khách hàng thành các nhóm (ví dụ: 3 cụm), từ đó phân loại khách theo hành vi.
6. **Dự đoán điểm tiêu dùng (Spending Score)**
 - Mức chi tiêu dự đoán của khách hàng mới giúp đánh giá **tiềm năng**: nếu cao hơn một ngưỡng nhất định (ví dụ 60) → được xếp vào nhóm khách hàng tiềm năng.

Phân nhóm khách hàng (3D)

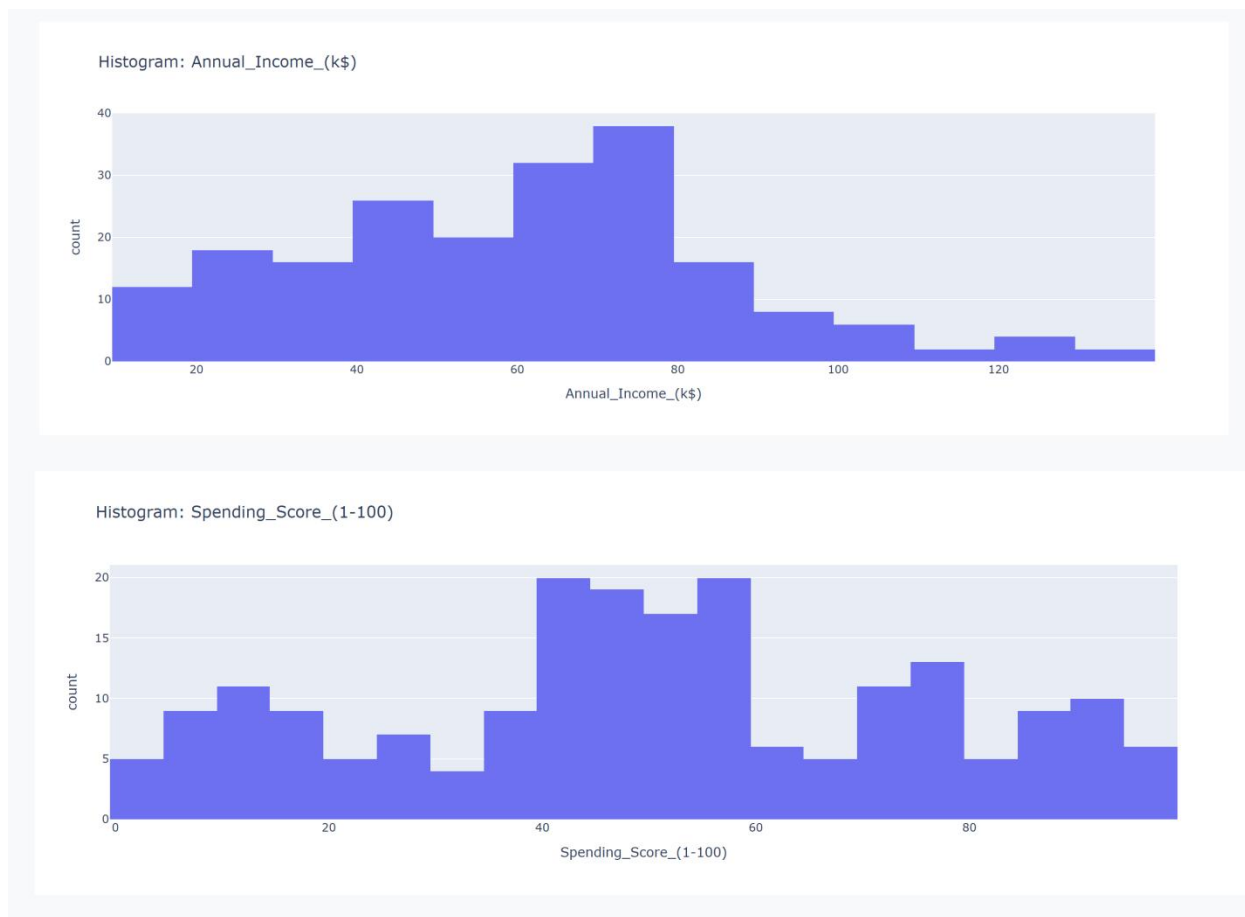


Scatter: Thu nhập vs Điểm tiêu dùng



Hình 8: Phân cụm KMeans





Hình 9: Biểu đồ Histogram

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Trong đề tài “*Phân loại khách hàng tiềm năng dựa trên dữ liệu tiêu dùng*”, nhóm đã xây dựng thành công một hệ thống web ứng dụng cho phép:

- **Phân cụm khách hàng** dựa trên các đặc trưng như tuổi, thu nhập, điểm tiêu dùng bằng thuật toán KMeans.
- **Dự đoán mức chi tiêu (Spending Score)** của khách hàng mới với mô hình học máy Random Forest Regression.
- **Hiển thị trực quan dữ liệu** bằng biểu đồ Histogram, Scatter Plot, và biểu đồ phân cụm 3D giúp người dùng dễ dàng nhận diện nhóm khách hàng tiềm năng.
- **Cho phép người dùng tải dữ liệu phân cụm** xuống dưới dạng file CSV để sử dụng cho các phân tích tiếp theo.
- **Xây dựng giao diện web hiện đại, trực quan, dễ sử dụng** bằng Flask và Bootstrap.

Hệ thống vận hành tốt, xử lý đúng dữ liệu và trả về kết quả chính xác, đáp ứng yêu cầu bài toán.

5.2. Hướng phát triển

Trong tương lai, hệ thống có thể mở rộng với các hướng sau:

- **Thêm các chỉ số khác** như: tần suất mua hàng, loại sản phẩm ưa thích, số lần tương tác với dịch vụ, v.v.
- **Áp dụng các thuật toán học sâu (Deep Learning)** cho dự đoán tiêu dùng chính xác hơn.
- **Xây dựng hệ thống gợi ý sản phẩm cá nhân hóa** dựa trên phân cụm khách hàng.
- **Kết nối hệ thống với cơ sở dữ liệu trực tiếp** thay vì file CSV để triển khai trong môi trường thực tế.
- **Tích hợp hệ thống phân tích theo thời gian thực**, xử lý khách hàng ngay khi có dữ liệu mới.

TÀI LIỆU THAM KHẢO

- Sử dụng AI: <https://chatgpt.com/>
- Trang web: [Data Science Tutorial](#)