**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**--------\*\*\*--------**

A blue button with white text

AI-generated content may be incorrect.

BÁO CÁO THỰC HÀNH

MÔN PHÂN TÍCH DỮ LIỆU

**TÊN ĐỀ TÀI PHÂN TÍCH KHÁM PHÁ DỮ LIỆU VỚI PYTHON**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | **ThS. Đỗ Như Tài** |
| **Nhóm sinh viên thực hiện:** |  |
| Hồ Thị Thanh Thảo | 3122410389 |
| Nguyễn Thị Hồng Thắm | 3122410392 |
| Phan Văn Thảo | 3122410391 |
| Nguyễn Hoàng Thiên Bảo | 3122410019 |

***Thành phố Hồ Chí Minh, tháng 09 năm 2025***

**LỜI MỞ ĐẦU**

Trong thời đại công nghệ 4.0, dữ liệu đã trở thành tài nguyên quý giá và đóng vai trò then chốt trong việc ra quyết định của các tổ chức và doanh nghiệp. Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA) là bước đầu tiên và quan trọng nhất trong quy trình khai thác dữ liệu, giúp chúng ta hiểu rõ bản chất, cấu trúc và các đặc điểm tiềm ẩn của tập dữ liệu trước khi tiến hành xây dựng mô hình học máy hay đưa ra các quyết định chiến lược.

Nhận thức được tầm quan trọng của EDA trong lĩnh vực khoa học dữ liệu, bài thực hành này tập trung vào việc trang bị cho sinh viên các kỹ năng cơ bản và cần thiết để thực hiện phân tích khám phá dữ liệu một cách có hệ thống và hiệu quả. Thông qua việc thực hành trên các tập dữ liệu thực tế, sinh viên sẽ được làm quen với các kỹ thuật thống kê mô tả, các phương pháp trực quan hóa dữ liệu, cũng như cách sử dụng các công cụ tự động hỗ trợ EDA.

Báo cáo này trình bày kết quả thực hành của chúng em về phân tích khám phá dữ liệu, bao gồm ba phần chính: Thống kê mô tả, Xử lý và trực quan hóa dữ liệu, và Phân tích đơn biến cùng hai biến. Mỗi phần không chỉ bao gồm kiến thức lý thuyết nền tảng mà còn có các bài tập thực hành cụ thể với code Python và phân tích kết quả chi tiết.

Qua quá trình thực hiện bài lab này, chúng em đã có cơ hội được áp dụng kiến thức đã học vào thực tế, từ đó hiểu sâu hơn về tầm quan trọng của việc khám phá và làm sạch dữ liệu trước khi xây dựng mô hình. Đồng thời, em cũng nắm vững cách sử dụng các thư viện Python phổ biến như Pandas, Matplotlib, Seaborn và các công cụ EDA tự động như SweetViz, AutoViz.

Em mong rằng báo cáo này sẽ thể hiện được sự nỗ lực học tập và nghiên cứu của em trong quá trình thực hành, đồng thời góp phần củng cố kiến thức nền tảng cho các môn học tiếp theo.

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn chân thành đến *Ths. Đỗ Như Tài*, người đã tận tình hướng dẫn và chia sẻ những kiến thức quý báu trong suốt quá trình thực hiện đề tài này. Sự chỉ bảo tận tình và những góp ý xây dựng từ thầy đã giúp chúng em hiểu rõ hơn về các vấn đề chuyên môn và hoàn thiện sản phẩm một cách tốt nhất.

Mặc dù đã rất cố gắng, nhưng do hạn chế về thời gian và kinh nghiệm, chắc chắn rằng không thể tránh khỏi những sai sót. Chúng em rất mong nhận được ý kiến đóng góp từ thầy để có thể cải thiện và hoàn thiện hơn nữa.

Một lần nữa, chúng em xin chân thành cảm ơn!

# PHẦN 1: GIỚI THIỆU VÀ MỤC TIÊU

## 1.1. Giới thiệu

### 1.1.1. Bối cảnh nghiên cứu

Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA) là quá trình phân tích tập dữ liệu để tóm tắt các đặc điểm chính của chúng, thường sử dụng các phương pháp trực quan hóa. EDA được John Tukey phát triển vào những năm 1970 và hiện nay đã trở thành bước không thể thiếu trong quy trình khoa học dữ liệu.

Trong thực tế, trước khi xây dựng bất kỳ mô hình máy học nào, việc hiểu rõ dữ liệu là vô cùng quan trọng. EDA giúp chúng ta:

* Phát hiện các vấn đề về chất lượng dữ liệu (missing values, outliers, duplicates)
* Hiểu phân bố và mối quan hệ giữa các biến
* Xác định các biến quan trọng cho việc modeling
* Đưa ra giả thuyết ban đầu về dữ liệu

### 1.1.2. Tầm quan trọng của EDA

Theo một nghiên cứu của Forbes, các data scientist dành khoảng 80% thời gian cho việc làm sạch và chuẩn bị dữ liệu, trong đó EDA chiếm phần lớn. Một EDA tốt có thể:

* Tiết kiệm thời gian và chi phí cho các bước tiếp theo
* Nâng cao độ chính xác của mô hình
* Giúp đưa ra quyết định đúng đắn về việc chọn thuật toán phù hợp

### 1.1.3. Công cụ và kỹ thuật

Trong bài thực hành này, chúng em sử dụng ngôn ngữ Python cùng với các thư viện phổ biến:

* **Pandas**: Xử lý và thao tác dữ liệu
* **NumPy**: Tính toán thống kê
* **Matplotlib**: Tạo biểu đồ cơ bản
* **Seaborn**: Trực quan hóa thống kê nâng cao
* **SweetViz**: Công cụ EDA tự động
* **AutoViz**: Công cụ visualization tự động

## 1.2. Mục tiêu bài thực hành

### 1.2.1. Mục tiêu chung

Bài thực hành này nhằm giúp sinh viên nắm vững các kỹ thuật cơ bản trong khám phá dữ liệu để hiểu rõ đặc điểm và cấu trúc của tập dữ liệu, từ đó đặt nền tảng cho các bước phân tích sâu hơn hoặc xây dựng mô hình khai thác dữ liệu.

### 1.2.2. Mục tiêu cụ thể

**Về kiến thức:**

* Hiểu và phân biệt được các khái niệm thống kê cơ bản (mean, median, mode, variance, standard deviation, quartiles, IQR)
* Nắm vững cách đọc và diễn giải các loại biểu đồ phổ biến (histogram, boxplot, scatter plot, heatmap)
* Phân biệt được phân tích đơn biến và phân tích hai biến
* Hiểu được tầm quan trọng của việc xử lý missing values và outliers

**Về kỹ năng:**

* Thực hiện thống kê mô tả trên tập dữ liệu thực tế
* Sử dụng thành thạo các thư viện Python: Pandas, Matplotlib, Seaborn
* Tạo các biểu đồ trực quan hóa dữ liệu hiệu quả
* Phát hiện và xử lý các vấn đề về chất lượng dữ liệu
* Sử dụng các công cụ EDA tự động (SweetViz, AutoViz)
* Phân tích mối quan hệ giữa các biến

**Về thái độ:**

* Có tư duy phân tích và khám phá dữ liệu có hệ thống
* Biết đặt câu hỏi đúng đắn về dữ liệu
* Cẩn thận, tỉ mỉ trong việc phân tích và đưa ra kết luận

### 1.2.3. Tập dữ liệu sử dụng

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên dataset** | **Nguồn** | **Mục đích** |
| 1 | Red Wine Quality | Kaggle | Thống kê mô tả & Trực quan hóa |
| 2 | Pima Indians Diabetes | Kaggle | Thống kê mô tả & Trực quan hóa |
| 3 | Marketing Campaign | Kaggle | Công cụ EDA tự động |

### 1.2.4. Nội dung thực hiện

Báo cáo được chia thành 3 phần chính:

**Phần 1: Thống kê mô tả**

* Ôn tập lý thuyết về các thước đo thống kê
* Thực hành tính toán các chỉ số thống kê
* Xử lý missing values và outliers

**Phần 2: Xử lý và trực quan hóa dữ liệu**

* Ôn tập lý thuyết về các loại biểu đồ
* Sử dụng Matplotlib và Seaborn
* Tạo các visualization có ý nghĩa

**Phần 3: Phân tích đơn biến và hai biến**

* Phân tích từng biến riêng lẻ
* Phân tích mối quan hệ giữa các biến
* Sử dụng công cụ EDA tự động

# PHẦN 2: THỐNG KÊ MÔ TẢ

## 2.1. Lý thuyết

**Câu hỏi 1: Thống kê mô tả là gì? Khác gì với thống kê suy luận?**

**Thống kê mô tả (Descriptive Statistics):**

* Là các phương pháp thu thập, tổ chức, trình bày và tóm tắt dữ liệu
* Mục đích: Mô tả đặc điểm cơ bản của tập dữ liệu hiện có
* Ví dụ: Tính trung bình, độ lệch chuẩn, vẽ biểu đồ

**Thống kê suy luận (Inferential Statistics):**

* Là phương pháp sử dụng dữ liệu mẫu để rút ra kết luận về tổng thể
* Mục đích: Dự đoán, kiểm định giả thuyết, ước lượng tham số
* Ví dụ: Kiểm định t-test, phân tích hồi quy, khoảng tin cậy

**Câu hỏi 2: Các thước đo thống kê mô tả chính (ví dụ: trung bình, trung vị, phương sai, độ lệch chuẩn) được sử dụng để làm gì? Trong trường hợp nào thì nên dùng trung vị thay vì trung bình?**

**Thước đo xu hướng trung tâm:**

* **Mean (Trung bình):** Tổng các giá trị chia cho số lượng quan sát
* **Median (Trung vị):** Giá trị ở giữa khi sắp xếp dữ liệu
* **Mode (Yếu vị):** Giá trị xuất hiện nhiều nhất

**Thước đo độ phân tán:**

* **Variance (Phương sai):** Đo lường mức độ phân tán quanh giá trị trung bình
* **Standard Deviation (Độ lệch chuẩn):** Căn bậc hai của phương sai

**Khi nào dùng Median thay vì Mean?**

* Khi dữ liệu có nhiều giá trị ngoại lai (outliers)
* Khi phân bố dữ liệu lệch (skewed distribution)
* Khi làm việc với dữ liệu thu nhập, giá nhà (thường có outliers)

**Câu hỏi 3: Làm thế nào để xác định phân bố của một tập dữ liệu? Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)?**

**Các loại phân bố phổ biến:**

* **Phân bố chuẩn (Normal):** Đối xứng, hình chuông
* **Phân bố lệch phải (Right-skewed):** Đuôi dài bên phải, mean > median
* **Phân bố lệch trái (Left-skewed):** Đuôi dài bên trái, mean < median

**Cách xác định:**

* Sử dụng histogram để quan sát hình dạng
* So sánh mean, median, mode
* Tính hệ số skewness

**Câu hỏi 4: Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu?**

**Độ lệch chuẩn (Standard Deviation):**

* Đo mức độ phân tán của dữ liệu xung quanh giá trị trung bình
* Giá trị nhỏ: Dữ liệu tập trung gần mean
* Giá trị lớn: Dữ liệu phân tán rộng

**Phạm vi (Range):**

* Range = Max - Min
* Cho biết khoảng biến thiên của dữ liệu
* Dễ bị ảnh hưởng bởi outliers

**Câu hỏi 5: Sự khác biệt giữa các thước đo như Q1, Q2, Q3 trong biểu đồ hộp (boxplot) là gì?**

**Các thước đo:**

* **Q1 (25th percentile):** 25% dữ liệu nhỏ hơn giá trị này
* **Q2 (Median, 50th percentile):** Trung vị
* **Q3 (75th percentile):** 75% dữ liệu nhỏ hơn giá trị này
* **IQR (Interquartile Range):** Q3 - Q1, đo mức độ phân tán của 50% dữ liệu giữa

**Xác định outliers:**

* Lower bound: Q1 - 1.5 × IQR
* Upper bound: Q3 + 1.5 × IQR

**Câu hỏi 6: Làm thế nào để xử lý giá trị thiếu (missing values) trước khi tính toán các chỉ số thống kê mô tả?**

**Phương pháp:**

1. **Loại bỏ:** Xóa các dòng/cột có giá trị thiếu (nếu số lượng nhỏ)
2. **Điền giá trị:** Mean, median, mode hoặc giá trị cụ thể
3. **Dự đoán:** Sử dụng các thuật toán ML để dự đoán giá trị thiếu
4. **Đánh dấu riêng:** Tạo category "Missing" cho biến phân loại

**Câu hỏi 7: Bạn có thể giải thích cách đọc và diễn giải một biểu đồ histogram hoặc boxplot từ dữ liệu thực tế không?**

**Histogram:**

* Trục X: Các khoảng giá trị (bins)
* Trục Y: Tần số xuất hiện
* Quan sát: Hình dạng phân bố, giá trị tập trung, độ phân tán

**Boxplot:**

* Hộp: Từ Q1 đến Q3
* Đường giữa hộp: Median
* Râu (whiskers): Kéo dài đến min/max trong khoảng hợp lệ
* Điểm ngoài râu: Outliers

**Câu hỏi 8: Khi gặp một tập dữ liệu có giá trị ngoại lai (outliers), bạn sẽ xử lý chúng như thế nào trước khi thực hiện thống kê mô tả?**

**Các phương pháp:**

1. **Loại bỏ:** Nếu là lỗi nhập liệu
2. **Giữ nguyên:** Nếu là giá trị hợp lệ quan trọng
3. **Biến đổi:** Log transformation, square root
4. **Winsorization:** Thay thế outliers bằng giá trị gần nhất trong ngưỡng
5. **Binning:** Gom nhóm dữ liệu

## 2.2. Bài tập thực hành