TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN NGÀNH**

**ĐỀ TÀI: GIẢM CHIỀU DỮ LIỆU BẰNG**

**PHƯƠNG PHÁP PCA**

**GVHD: TS. Lê Quang Minh**

**LỚP: DH21CS02**

**Ngành: Khoa Học Máy Tính**

**SINH VIÊN THỰC HIỆN:**

**Huỳnh Nguyễn Bảo Trân - 2151010390**

**Thành phố Hồ Chí Minh, 2024**

# LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành đến thầy Lê Quang Minh, người đã tận tình hướng dẫn và luôn sẵn sàng hỗ trợ em trong suốt quá trình thực hiện đồ án. Nhờ sự chỉ dẫn chi tiết, cùng những lời khuyên, những kinh nghiệm, kiến thức thầy truyền lại, em đã hoàn thành đồ án này.

Em cũng xin bày tỏ lòng biết ơn đến Trường Đại học Mở, cùng các thầy cô trong Khoa Công nghệ thông tin, đã chỉ dạy cho em những kiến thức nền tảng và tạo điều kiện thuận lợi cho em trong suốt quá trình học tập và làm đồ án.

Cuối cùng, em xin chân thành cảm ơn toàn thể quý thầy cô của trường Đại học Mở đã luôn chia sẻ kiến thức, hỗ trợ và đồng hành cùng em trong suốt quá trình học tập tại trường.

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

# MỤC LỤC

[LỜI CẢM ƠN 2](#_Toc178953753)

[DANH MỤC HÌNH ẢNH 6](#_Toc178953755)

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 7](#_Toc178953756)

[**1.1** **Lý do chọn đề tài** 7](#_Toc178953757)

[**1.2** **Mục tiêu nghiên cứu** 8](#_Toc178953758)

[**1.3** **Phạm vi nghiên cứu** 8](#_Toc178953759)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 9](#_Toc178953760)

[**2.1** **GIỚI THIỆU VỀ GIẢM CHIỀU DỮ LIỆU** 9](#_Toc178953761)

[**2.1.1** **Khái niệm cơ bản về giảm chiều dữ liệu** 9](#_Toc178953762)

[**2.1.2** **Các phương pháp giảm chiều dữ liệu phổ biến** 9](#_Toc178953763)

[**2.2** **GIỚI THIỆU VỀ PHƯƠNG PHÁP PCA** 9](#_Toc178953764)

[**2.2.1** **Khái niệm cơ bản về phương pháp PCA** 9](#_Toc178953765)

[**2.2.2** **Nguyên lý hoạt động của phương pháp PCA** 10](#_Toc178953766)

[**2.3** **MÔI TRƯỜNG VÀ NGÔN NGỮ LẬP TRÌNH** 10](#_Toc178953767)

[**2.3.1** **Google Colab** 10](#_Toc178953768)

[**2.3.2** **Python** 11](#_Toc178953769)

[CHƯƠNG 3: THỰC NGHIỆM 12](#_Toc178953770)

[**3.1** **MÔ TẢ BỘ DỮ LIỆU** 12](#_Toc178953771)

[**3.2** **QUY TRÌNH TIỀN XỬ LÝ DỮ LIỆU** 12](#_Toc178953772)

[**3.3** **MỐI QUAN HỆ TƯƠNG QUAN** 17](#_Toc178953773)

[**3.4** **ỨNG DỤNG PCA** 18](#_Toc178953774)

[**3.5** **ĐÁNH GIÁ KẾT QUẢ** 21](#_Toc178953775)

[CHƯƠNG 4: KẾT LUẬN 22](#_Toc178953776)

[CHƯƠNG 5: TÀI LIỆU THAM KHẢO 23](#_Toc178953777)

**DANH MỤC HÌNH ẢNH**

[Hình 3. 1: Code đọc tệp và đường dẫn tới file dữ liệu 11](#_Toc178867105)

[Hình 3. 2: Dữ liệu thô ban đầu, chưa qua xử lý 12](#_Toc178867106)

[Hình 3. 3: Code kiểm tra và kết quả dữ liệu trùng lập 12](#_Toc178867107)

[Hình 3. 4: Kiểm tra dữ liệu trống 13](file:///C:\Đồ%20Án%20Ngành\Báo%20cáo%20đồ%20án%20ngành.docx#_Toc178867108)

[Hình 3. 5: Kiểm tra dữ liệu trống lần 2 14](file:///C:\Đồ%20Án%20Ngành\Báo%20cáo%20đồ%20án%20ngành.docx#_Toc178867109)

[Hình 3. 6: Kết quả sau khi xử lý dữ liệu trống 15](file:///C:\Đồ%20Án%20Ngành\Báo%20cáo%20đồ%20án%20ngành.docx#_Toc178867110)

[Hình 3. 7: Code mã hóa dữ liệu cột Sex và Embarked 15](#_Toc178867111)

[Hình 3. 8: Code loại bỏ cột Name, Ticket và PassengerId 16](#_Toc178867112)

[Hình 3. 9: Code chuẩn hóa dữ liệu 16](#_Toc178867113)

[Hình 3. 10: Biểu đồ biểu diễn mối quan hệ giữa các thuộc tính 16](#_Toc178867114)

[Hình 3. 11: Code ứng dụng PCA lên dữ liệu đã chuẩn hóa 17](#_Toc178867115)

[Hình 3. 12: Biểu đồ thể hiện tỷ lệ phương sai được giải thích 17](#_Toc178867116)

[Hình 3. 13: Code so sánh phương sai được giải thích 18](#_Toc178867117)

[Hình 3. 14: Kết quả thu được sau khi so sánh phương sai 18](#_Toc178867118)

[Hình 3. 15: Biểu đồ biểu diễn ma trận tải 19](#_Toc178867119)

[Hình 3. 16: Code thực hiện PCA và kết quả số lượng thành phần được giữ lại 20](file:///C:\Đồ%20Án%20Ngành\Báo%20cáo%20đồ%20án%20ngành.docx#_Toc178867120)

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

## **Lý do chọn đề tài**

Hiện nay, trong thời đại chuyển đổi số một cách mạnh mẽ thì việc lưu trữ dữ liệu về hình ảnh, thông tin, âm thanh……ngày càng nhiều. Từ đó các dữ liệu lớn cũng ngày càng phát triển nhằm phục vụ cho việc ứng dụng vào thực tiễn như đưa ra kết quả quyết định, dự báo, nhận dạng mẫu dựa trên kết quả phân tích dữ liệu. Việc xử lý và phân tích dữ liệu có thể được ứng dụng trong nhiều lĩnh vực khác nhau như kinh tế, chính trị, giáo dục, y tế và nhiều lĩnh vực khác.

Có thể thấy các dữ liệu ngày càng lớn cũng đã trở thành thách thức đối với những người làm trong ngành phân tích dữ liệu. Việc xử lý và phân tích một bộ dữ liệu lớn phức tạp và nhiều chiều có thể gây khó khăn trong việc đưa ra kết quả. Vì khi dữ liệu có quá nhiều thuộc tính (features), nó có thể dẫn đến việc kết quả phân tích bị sai lệch do có quá nhiều dữ liệu, thông tin gây nhiễu. Đồng thời, nó cũng làm cho việc xử lý, phân tích và xây dựng mô hình học máy trở nên khó khăn và tốn kém tài nguyên. Nên để đưa ra được kết quả phân tích có độ chính xác cao thì việc giảm chiều dữ liệu mà vẫn giữ được phần lớn các thông tin quan trọng là rất cần thiết.

Phương pháp Principal Component Analysis (PCA) là một phương pháp giảm chiều dữ liệu phổ biến và hiệu quả. Đây là phương pháp giúp đơn giản hóa quá trình phân tích và xây dựng mô hình máy học trên các bộ dữ liệu lớn. PCA có thể chuyển đổi một bộ dữ liệu lớn thành một bộ dữ liệu nhỏ hơn nhưng vẫn giữ được phần lớn thông tin quan trọng. Ngoài ra, PCA còn giúp loại bỏ các dữ liệu gây nhiễu và không quan trọng. Điều này sẽ làm cho kết quả phân tích được đưa ra có độ chính xác cao, giảm thời gian tính toán và thời gian xây dựng mô hình.

Do đó, em đã chọn đề tài “Giảm chiều dữ liệu bằng phương pháp PCA” để nghiên cứu nguyên lý và tính ứng dụng của phương pháp này.

## **Mục tiêu nghiên cứu**

Mục tiêu mà em muốn nghiên cứu trong đề tài này là áp dụng giảm chiều dữ liệu thành công bằng phương pháp PCA trên bộ dữ liệu Titanic. Kết quả em mong muốn là xác định được số lượng thành phần chính cần thiết giữ lại có thể giải thích được tối thiểu 90% phương sai của dữ liệu gốc. Thành công thu gọn dữ liệu mà vẫn giữ được phần lớn thông tin quan trọng. Cuối cùng là đánh giá được hiệu suất của phương pháp PCA.

## **Phạm vi nghiên cứu**

Phạm vi nghiên cứu của đề tài sẽ tập trung vào việc tìm hiểu và áp dụng phương pháp PCA để giảm chiều dữ liệu. Nghiên cứu sẽ được thực hiện trên bộ dữ liệu Titanic, chứa các thông tin về hành khách trên con tàu Titanic. Phương pháp PCA sẽ được áp dụng để xác định số lượng thành phần cần giữ lại, giảm chiều của bộ dữ liệu Titanic và giữ lại được các thông tin quan trọng từ bộ dữ liệu gốc. Hiệu suất của phương pháp sẽ được đánh giá thông qua tỷ lệ phương sai được giải thích của các thành phần chính cần giữ lại. Trong quá trình nghiên cứu đề tài sẽ dùng ngôn ngữ Python và môi trường Google Colab để thực hiện nghiên cứu.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## **GIỚI THIỆU VỀ GIẢM CHIỀU DỮ LIỆU**

### **Khái niệm cơ bản về giảm chiều dữ liệu**

Giảm chiều dữ liệu (Dimensionality Reduction) là quá trình biến đổi dữ liệu từ không gian có số chiều cao về không gian có số chiều thấp, nhưng vẫn giữ được những thông tin quan trọng của dữ liệu [1]. Giảm chiều dữ liệu sẽ làm cho bộ dữ liệu nhiều thuộc tính thành bộ dữ liệu ít thuộc tính hơn. Làm cho việc xử lý, phân tích và xây dựng mô hình máy học dễ dàng hơn, ít tốn tài nguyên và thời gian hơn. Kết quả đưa ra từ việc phân tích bộ dữ liệu có độ chính xác và tính ứng dụng cao.

### **Các phương pháp giảm chiều dữ liệu phổ biến**

Phương pháp giảm chiều dữ liệu được chia thành 2 hướng là lựa chọn thuộc tính và trích xuất thuộc tính.

Giảm chiều dữ liệu theo hướng lựa chọn thuộc tính gồm các phương pháp: Missing values ratio, Low-variance filter, High-correlation filter, Random forest, …..

Giảm chiều dữ liệu theo hướng trích xuất thuộc tính gồm các phương pháp: Factor analysis, PCA, Linear Discriminatory Analysis, Singular Value Composition,…. [2].

## **GIỚI THIỆU VỀ PHƯƠNG PHÁP PCA**

### **Khái niệm cơ bản về phương pháp PCA**

PCA là phương pháp tuyến tính thông dụng nhất để giảm chiều dữ liệu. PCA thực hiện một phép biến đổi tuyến tính, chuyển dữ liệu sang không gian thấp chiều hơn [1]. Đây là một phương pháp phổ biến giúp chuyển đổi tập dữ liệu nhiều thuộc tính thành tập dữ liệu ít thuộc tính hơn nhưng vẫn giữ được thông tin quan trọng từ tập dữ liệu gốc. Ngoài ra, PCA còn giúp cải thiện hiệu suất của mô hình học máy, loại bỏ nhiễu trong dữ liệu, trực quan hóa dữ liệu trong không gian thấp chiều.

### **Nguyên lý hoạt động của phương pháp PCA**

PCA hoạt động bằng cách tìm các vector riêng và giá trị riêng của ma trận hiệp phương sai của dữ liệu gốc. Các vector riêng là các trục chính và các giá trị riêng cho biết mức độ biến đổi trong dữ liệu được mỗi thành phần chính thu được [3].

Các bước chính của phương pháp PCA:

* Chuẩn hóa dữ liệu: Chuyển đổi các giá trị của các thuộc tính về cùng một thang đo.
* Tính ma trận hiệp phương sai: Tính ma trận hiệp phương sai của dữ liệu đã chuẩn hóa để xác định các mối quan hệ tương quan giữa các thuộc tính.
* Tìm vector riêng và giá trị riêng: Tìm các thành phần chính mới có phương sai lớn chứa nhiều thông tin trong bộ dữ liệu.
* Sắp xếp các thành phần chính: Lựa chọn các thành phần chính mang nhiều thông tin có ý nghĩa nhất. Sắp xếp theo chiều giảm dần.
* Giảm số chiều: Tạo tập dữ liệu mới với số lượng thành phần chính cần thiết giữ lại. Dữ liệu đã được giảm số chiều nhưng vẫn giữ các thông tin quan trọng của tập dữ liệu gốc.

1. **MÔI TRƯỜNG VÀ NGÔN NGỮ LẬP TRÌNH**
2. **Google Colab**

Google Colab là một dịch vụ miễn phí từ Google cho phép người dùng viết và chia sẻ mã Python [4]. Người dùng có thể chạy mã Python trực tiếp trên trình duyệt mà không cần tải bất kỳ ứng dụng nào về máy.

Google Colab được thiết kế hỗ trợ học máy và khoa học dữ liệu, có tích hợp google drive, nhiều thư viện Python phổ biến. Vì vậy, đối với những người làm trong ngành phân tích dữ liệu và đặc biệt là sinh viên ngành khoa học dữ liệu thì đây là một công cụ hỗ trợ công việc, học tập dễ sử dụng, hữu ích và miễn phí.

1. **Python**

Python là ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học [5]. Python được các lập trình viên sử dụng rất phổ biến và rộng rãi bởi vì ngôn ngữ có cú pháp đơn giản, dễ học. Ngoài ra, Python còn hoạt động được trên nhiều nền tảng khác nhau và có nhiều thư viện hỗ trợ. Do đó, đây là một ngôn ngữ lý tưởng và phù hợp với cả người mới bắt đầu học lập trình và lập trình viên lâu năm.

# CHƯƠNG 3: THỰC NGHIỆM

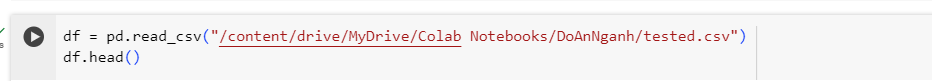
## **MÔ TẢ BỘ DỮ LIỆU**

Bộ dữ liệu Titanic chứa số liệu, thông tin của các hành khách trên con tàu Titanic trong vụ chìm tàu nổi tiếng xảy ra vào năm 1912, bao gồm:

1. Mã định danh của mỗi hành khách (PassengerId)
2. Kết quả sống sót (Survived)
3. Hạng vé của hành khách (Pclass)
4. Tên của hành khách (Name)
5. Giới tính (Sex)
6. Tuổi (Age)
7. Số lượng anh chị em hoặc vợ chồng đi cùng (SibSp)
8. Số lượng cha mẹ hoặc con đi cùng (Parch)
9. Số vé của hành khách (Ticket)
10. Giá vé của hành khách (Fare)
11. Số phòng của hành khách (Cabin)
12. Cảng mà hành khách lên tàu (Embarked)
13. **QUY TRÌNH TIỀN XỬ LÝ DỮ LIỆU**

Vì dữ liệu lớn nên em sẽ sử dụng thư viện pandas trong python để hỗ trợ. Đối với bộ dữ liệu Titanic em dùng môi trường của Google Colab để thực hiện tiền xử lý dữ liệu.

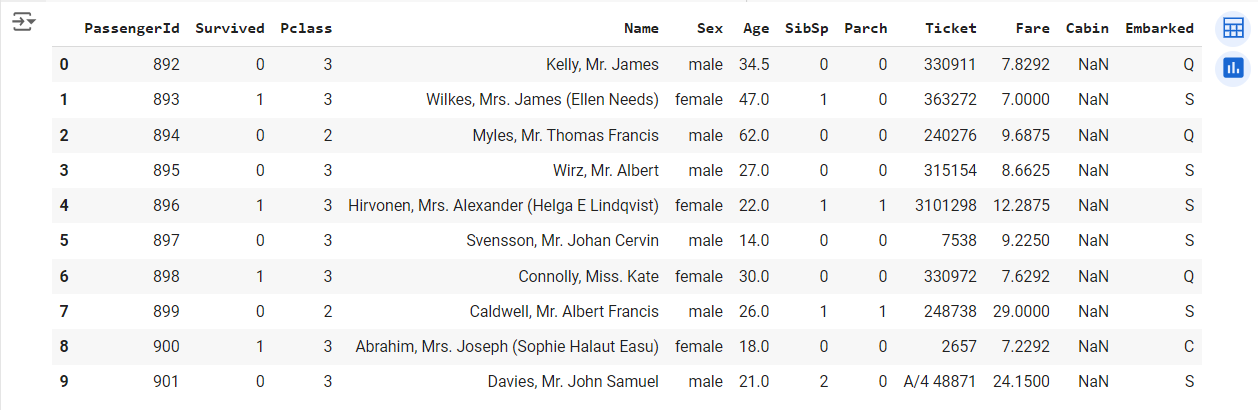
Đầu tiên, em xác định đường dẫn đến file dữ liệu và đọc tệp vào Pandas DataFrame bằng thư viện pandas.



Hình 3. 1: Code đọc tệp và đường dẫn tới file dữ liệu

Sau khi chạy xong lệnh đọc file dữ liệu, Notebook sẽ hiện dữ liệu gồm 418 dòng và 12 cột, lưu thông tin gồm 418 hành khách trên tàu Titanic vào năm 1912.

Dữ liệu của 10 dòng đầu trong tập dữ liệu Titanic.

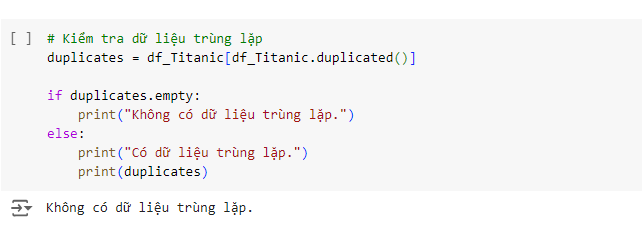


Hình 3. 2: Dữ liệu thô ban đầu, chưa qua xử lý

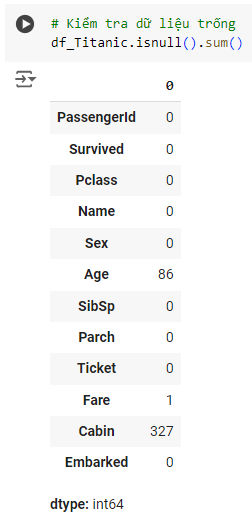
Chiều dài và kích thước của bộ dữ liệu

* Len: 418
* Total 12 columns
* Dtypes: float64(2), int64(5), object(5)

Để tránh kết quả thu được không có độ chính xác bởi các dữ liệu lặp gây nhiễu thì em đã kiểm tra dữ liệu trùng lặp của bộ dữ liệu Titanic. Kết quả cho thấy bộ dữ liệu không có dữ liệu lặp lại.

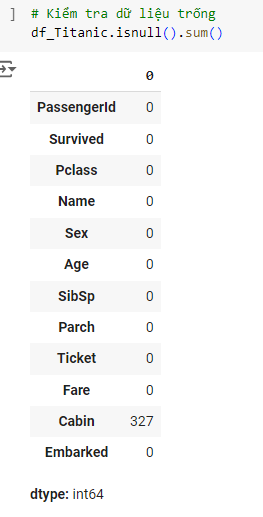


Hình 3. 3: Code kiểm tra và kết quả dữ liệu trùng lập

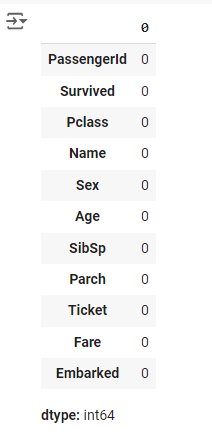
Tiếp theo, em đã kiểm tra dữ liệu bị thiếu trong bộ dữ liệu. Để đảm bảo rằng kết quả không bị sai lệch do thiếu dữ liệu thì việc xử lý các dữ liệu thiếu cũng rất quan trọng.

Hình 3. 4: Kiểm tra dữ liệu trống

Sau khi chạy lệnh kiểm tra dữ liệu trống, chúng ta sẽ biết được số lượng và các dữ liệu trong cột nào đang bị thiếu. Từ kết quả trên ta có thể thấy cột Age có 86 dòng dữ liệu trống, cột Fare có 1 dòng dữ liệu trống và cột Cabin có 327 dòng dữ liệu trống.

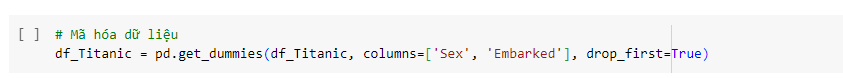
Để xử lý các dòng dữ liệu bị trống, em đã điền giá trị trung vị vào để lấp đầy dữ liệu bị thiếu. Sau khi điền giá trị cho các dữ liệu thiếu, em tiến hành kiểm tra dữ liệu trống thêm lần nữa.

Hình 3. 5: Kiểm tra dữ liệu trống lần 2

Kết quả vẫn còn cột Cabin vẫn còn dữ liệu trống. Vì số dòng dữ liệu trống của cột Cabin chiếm phần lớn trong bộ dữ liệu nên em xử lý dữ liệu trống bằng cách xóa cột Cabin.

Hình 3. 6: Kết quả sau khi xử lý dữ liệu trống

Sau khi xử lý xong các dữ liệu bị thiếu, em sử dụng One – Hot Encoding để mã hóa cột Sex và Embarked. Để thực hiện PCA lên bộ dữ liệu thì cần phải mã hóa các biến phân loại thành giá trị số.

**

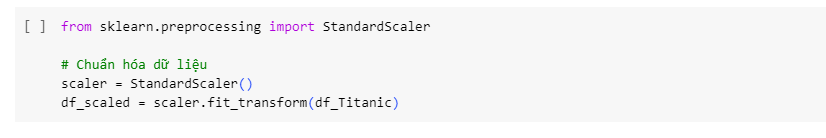
Hình 3. 7: Code mã hóa dữ liệu cột Sex và Embarked

Tiếp theo, em loại bỏ cột Name, Ticket và PassengerId. Vì PCA là phương pháp giảm chiều dữ liệu dựa trên mối quan hệ giữa các thuộc tính trong tập dữ liệu. Trong khi các cột dữ liệu này có mức độ cá nhân hóa cao và không có ý nghĩa trong việc phân tích PCA.



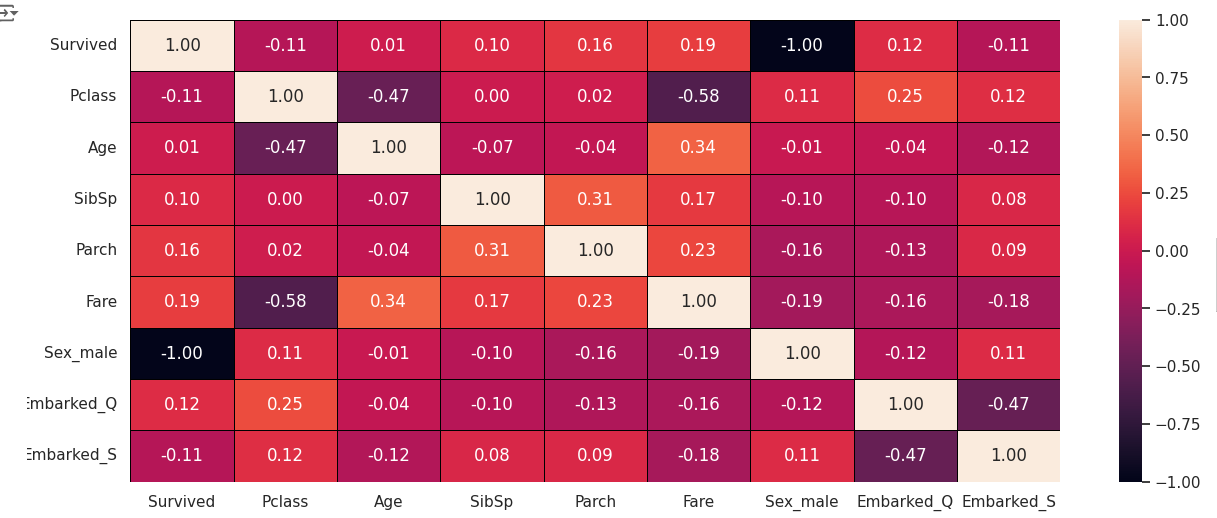
Hình 3. 8: Code loại bỏ cột Name, Ticket và PassengerId

Để thực hiện PCA thì các thuộc tính của bộ dữ liệu phải có cùng thang đo. Do đó trước khi thực hiện PCA thì phải chuẩn hóa bộ dữ liệu.



Hình 3. 9: Code chuẩn hóa dữ liệu

1. **MỐI QUAN HỆ TƯƠNG QUAN**

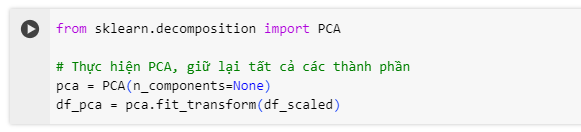


Hình 3. 10: Biểu đồ biểu diễn mối quan hệ giữa các thuộc tính

Chỉ số về mối quan hệ tương quan dao động từ [-1.00] đến [1.00]. Tổng quan ta thấy, đối với các cặp thuộc tính có mức tương quan gần 1, có mối quan hệ mạnh mẽ. Các cặp thuộc tính với mức tương quan âm gần -1 thường có mối quan hệ nghịch biến mạnh mẽ. Giá trị tương quan ở mức trung bình từ 0.75-0.50, có thể có mối quan hệ tương đối mạnh nhưng không đến mức rất mạnh. Khi giá trị tương quan gần 0, có thể cho thấy sự độc lập giữa các thuộc tính, không có mối quan hệ tuyến tính rõ ràng.

1. **ỨNG DỤNG PCA**

Sử dụng tập dữ liệu đã được chuẩn hóa, tiến hành ứng dụng PCA lên dữ liệu và giữ lại tất cả các thành phần.



Hình 3. 11: Code ứng dụng PCA lên dữ liệu đã chuẩn hóa

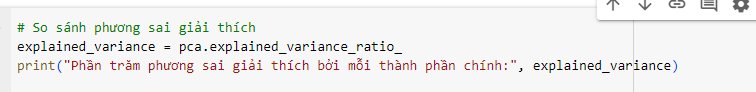
Tiếp theo, em đã vẽ biểu đồ về tỷ lệ phương sai được giải thích để biết được phần trăm dữ liệu gốc được giữ lại sau khi thực hiện PCA.



Hình 3. 12: Biểu đồ thể hiện tỷ lệ phương sai được giải thích

Thông qua biểu đồ Elbow trên, chúng ta có thể thấy đường cong đi lên một cách nhanh chóng chứng tỏ các thành phần chính từ PC1 đến PC4 đóng góp nhiều phương sai hơn. Nơi đường cong bắt đầu phẳng dần là từ PC5 đến PC9. Điều này cho biết rằng các thành phần chính sau PC5 không giải thích nhiều phương sai của dữ liệu.

Để xác định được số lượng thành phần chính cần giữ lại để giảm chiều dữ liệu mà vẫn giữ được phần lớn thông tin quan trọng, chúng ta cần so sánh phương sai được giải thích.



Hình 3. 13: Code so sánh phương sai được giải thích

Sau khi so sánh phương sai, em thu được kết quả:



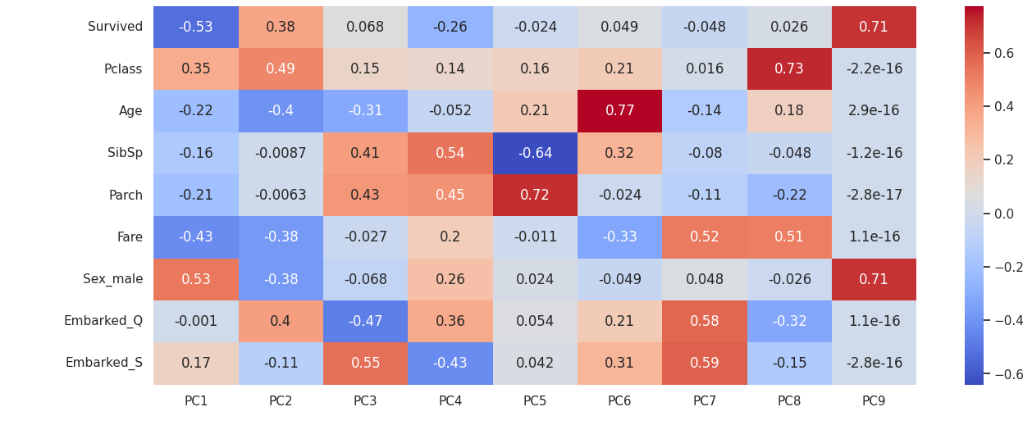
Hình 3. 14: Kết quả thu được sau khi so sánh phương sai

Từ kết quả trên cho ta biết phần trăm phương sai được giải thích bởi các thành phần chính như sau:

* + Thành phần chính PC1 giải thích 26.16% phương sai trong dữ liệu.
  + Thành phần chính PC2 giải thích 20.22% phương sai trong dữ liệu.
  + Thành phần chính PC3 giải thích 18.11% phương sai trong dữ liệu.
  + Thành phần chính PC4 giải thích 12.22% phương sai trong dữ liệu.
  + Thành phần chính PC5 giải thích 7.74% phương sai trong dữ liệu.
  + Thành phần chính PC6 giải thích 7.26% phương sai trong dữ liệu.
  + Thành phần chính PC7 giải thích 4.50% phương sai trong dữ liệu.

Càng về các thành phần chính cuối cùng phần trăm phương sai được giải thích càng thấp. Điều này cho thấy các thành phần chính cuối cùng có thể chứa các thông tin nhiễu hoặc không quan trọng. Vậy để giảm chiều dữ liệu thì chúng ta chỉ cần giữ lại 6 thành phần chính tương đương phần trăm phương sai được giải thích là 91.71%, gần như bảo toàn được toàn bộ thông tin quan trọng của dữ liệu gốc.

Tiếp theo, để biết được độ ảnh hưởng của thuộc tính đến các thành phần, ta cần phải kiểm tra ma trận tải. Dùng biểu đồ heatmap để biểu diễn ma trận tải:

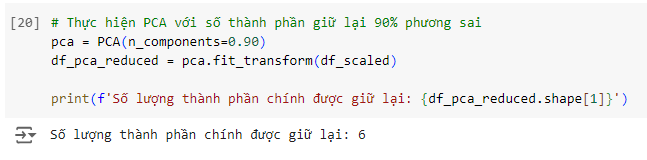


Hình 3. 15: Biểu đồ biểu diễn ma trận tải

Kết quả trên cho ta thấy được mối tương quan giữa các thuộc tính với các thành phần chính dựa trên số phương sai và màu sắc của các ô trong biểu đồ. Phương sai càng lớn (dương hoặc âm), màu sắc càng đậm cho thấy thuộc tính đó có ảnh hưởng rất đến thành phần chính tương ứng. Và giá trị dương (màu đỏ) có nghĩa là thuộc tính ảnh hưởng cùng chiều với thành phần chính, trong khi giá trị âm (màu xanh) cho biết thuộc tính ảnh hưởng ngược chiều.

Điển hình là thành phần chính PC1 có:

* Thuộc tính Survived (-0.53) có giá trị âm và màu sắc ô là xanh dương đậm, điều này cho thấy PC1 là đại diện cho sự sống sót, nhưng theo chiều ngược lại.
* Thuộc tính Sex\_male (0.53) có giá trị dương và màu sắc ô là cam đậm cho thấy thuộc tính giới tính nam có đóng góp mạnh vào PC1.
* Mối quan hệ của 2 thuộc tính trên đối với PC1 có thể được hiểu là tỷ lệ sống sót của hành khách nam trên tàu Titanic là rất thấp.

Sau khi xác định được số lượng phương sai cần thiết giữ lại chúng ta có thể giảm chiều dữ liệu bằng PCA thêm lần nữa để dữ liệu trở nên ít hơn mà vẫn giữ được lượng lớn dữ liệu quan trọng.

Hình 3. 16: Code thực hiện PCA và kết quả số lượng thành phần được giữ lại

1. **ĐÁNH GIÁ KẾT QUẢ**

Để đánh giá được hiệu suất của PCA thì ta cần so sánh dữ liệu trước và sau khi thực hiện PCA để giảm chiều dữ liệu. Em chọn so sánh dữ liệu trước và sau khi giảm chiều ảnh hưởng đến thuộc tính Survived như thế nào.

Thông qua biểu đồ heatmap biểu diễn mối quan hệ giữa các thuộc tính trước khi giảm chiều dữ liệu (hình 3.10). Biểu đồ cho thấy rõ ràng mối quan hệ giữa các thuộc tính như “Sex\_male” đặc biệt có mối quan hệ mật thiết và ảnh hưởng rất lớn đến thuộc tính Survived đánh giá tỷ lệ sống sót của hành khách, “Fare”, “Pclass” và “Age” có ảnh hưởng lớn đến khả năng sống sót.

Trong khi đó, biểu đồ heatmap biểu diễn ma trận tải (hình 3.15) cho thấy một số thuộc tính có mối tương quan cao với các thành phần chính. Đặc biệt là thành phần chính PC1 có mối tương quan cao với thuộc tính Survived đại diện cho sự sống sót của hành khách. Các thuộc tính có mối tương quan cao với PC1 như là “Sex\_male”, “Fare”, “Pclass” cũng giống như dữ liệu trước khi thực hiện PCA nhưng làm rõ các mối quan hệ chính hơn. Đây là những thuộc tính trực tiếp ảnh hưởng đến khả năng sống sót của hành khách.

Như vậy có thể thấy việc thực hiện PCA sẽ làm cho dữ liệu trở nên nhỏ hơn mà không làm mất đi thông tin quan trọng của dữ liệu gốc, loại bỏ được dữ liệu chứa thông tin gây nhiễu, giúp cải thiện hiệu suất và khả năng dự đoán của mô hình.

# CHƯƠNG 4: KẾT LUẬN

Trong đề tài này, em đã nghiên cứu giảm chiều dữ liệu bằng phương pháp PCA trên bộ dữ liệu Titanic nhằm giảm chiều của bộ dữ liệu gốc, đơn giản hóa quá trình phân tích, trực quan hóa dữ liệu. Phương pháp PCA được chọn bởi vì khả năng bảo toàn được phần lớn thông tin quan trọng của tệp dữ liệu gốc bằng việc giữ lại các thành phần chính quan trọng.

Qua quá trình thực nghiệm cho thấy, chỉ cần giữ lại 6 thành phần chính đã có thể giữ lại 90% phương sai của dữ liệu gốc, gần như bảo toàn được toàn bộ thông tin quan trọng của bộ dữ liệu gốc. Điều này chứng minh rằng việc giảm chiều dữ liệu bằng phương pháp PCA có thể loại bỏ những thông tin nhiễu, không quan trọng và giữ lại được những yếu tố quan trọng nhất, tăng độ chính xác cho mô hình học máy.

Em đã thực hiện giảm chiều dữ liệu bằng phương pháp PCA thành công trên bộ dữ liệu Titanic. Dựa vào kết quả thu được sau khi thực nghiệm cho thấy PCA là một phương pháp giảm chiều dữ liệu hiệu quả. PCA chỉ làm gọn dữ liệu mà không làm mất thông tin quan trọng của bộ dữ liệu gốc, giữ lại các thành phần chính và loại bỏ các thành phần không quan trọng. Giúp cho việc phân tích, trực quan hóa và xây dựng mô hình trở nên dễ dàng, đơn giản hơn.

# CHƯƠNG 5: TÀI LIỆU THAM KHẢO

[1] **“**Tổng quan về Giảm chiều dữ liệu và Principle Component Analysis”.[Trực tuyến]. Địa chỉ: <https://tek4.vn/khoa-hoc/machine-learning-co-ban/giam-chieu-du-lieu> [Truy cập 16/09/202]

[2] Blink, “Các kỹ thuật Dimensionality Reduction”, 05/07/2021. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/cac-ky-thuat-dimensionality-reduction-OeVKB98A5kW> [Truy cập 19/09/2024]

[3] Phan Ngoc, “Giới thiệu Principal Component Analysis”, 22/09/2019. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/gioi-thieu-principal-component-analysis-07LKXpq2KV4> [Truy cập 19/09/2024]

[4] Nhut Lien, “Google Colab là gì? Khám phá những tiện ích tuyệt vời và cách sử dụng Google Colab”, 9 tháng trước. [Trực tuyến]. Địa chỉ: <https://fptshop.com.vn/tin-tuc/danh-gia/google-colab-167087> [Truy cập 20/09/2024]

[5] “Python là gì?”. [Trực tuyến]. Địa chỉ: <https://aws.amazon.com/vi/what-is/python/> [Truy cập 20/09/2024]