TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: PHÂN CỤM MỨC ĐỘ ƯU TIÊN BỆNH NHÂN**

Giáo viên hướng dẫn: TS. Nguyễn Thị Kim Ngân

Sinh viên/nhóm sinh viên thực hiện:

1. Nguyễn Văn Xuân, Lớp 63CNTT2

2. Trương Quốc Nam, Lớp 63CNTT2

3. Nguyễn Thành Vũ, Lớp 63CNTT2

**Hà Nội, tháng 11 năm 2023**

**MỤC LỤC**

**Phần 1. Tổng quan**

***1. Giới thiệu về học máy***

- Lịch sử và vai trò của Machine Learning

* Lịch sử của Machine Learning (Học Máy): Là thuật ngữ được đặt bởi Arthur Samuel vào năm 1959. Samuel là một IBMer người Mỹ kiêm nhà tiên phong trong lĩnh vực trí tuệ nhân tạo và máy tính chơi game. Năm 1960, thuật ngữ học máy phố biển hơn thông qua cuốn sách của Nilsson, nội dung đề cập đến việc phân loại học máy. Machine Learning hiện đại bao gồm hai mục tiêu chính: phân loại dữ liệu thông qua mô hình đã được phát triển và đưa ra dự đoán về kết quả trong tương lai dựa trên mô hình này.
* Vai trò của Machine Learning (Học Máy):
* Sử dụng các nguyên tắc cơ bản của khoa học máy tính như thuật toán, cấu trúc dữ liệu, độ phức tạp và kiến trúc máy tính
* Sử dụng nhiều kỹ năng toán học để thực hiện các phép tính và tính toán cũng như làm việc với các thuật toán dựa trên lập trình
* Tạo ra kết quả dự án và tách biệt các vấn đề yêu cầu giải quyết để mang lại hiệu quả hơn trong các mã và chương trình
* Cộng tác với dữ liệu để phát triển các đường ống mô hình khác nhau
* Quản lý đường ống dữ liệu và cơ sở hạ tầng để tọa mã
* Tạo thuật toán trên cơ sở các mô hình và quy trình thống kê khác nhau
* Sử dụng các chiến lược đánh giá dữ liệu và mô hình hóa dữ liệu để xác định các mẫu và mô tả các trường hợp trong tương lai
* Áp dụng các thuật toán và thư viện Machine Learning có liên quan
* Phân tích tập hợp lớn và phức tạp của dữ liệu có cấu trúc và không có cấu trúc để có được thông tin chi tiết và quyết định các kỹ thuật
* Nghiên cứu và triển khai thêm các kỹ thuật và thực tiễn mới nhất để cải thiện cơ sở hạ tầng của tổ chức

- Ưu điểm và hạn chế của Học có giám sát và không giám sát

* Học có giám sát
* Ưu điểm:
* Học tập có giám sát cho phép thu thập dữ liệu và tạo ra dữ liệu đầu ra từ những kinh nghiệm trước đó.
* Giúp tối ưu hóa các tiêu chí hiệu suất với sự trợ giúp của kinh nghiệm
* ML có giám sát giúp giải quyết nhiều loại vấn đề tính toán trong thế giới thực
* Nhược điểm:
* Phân loại dữ liệu lớn có thể là một thách thức.
* Đào tạo cho việc học có giám sát cần rất nhiều thời gian tính toán, vì vậy, nó đòi hỏi rất nhiều thời gian.
* Học không giám sát:
* Ưu điểm:
* Không cần phải có dữ liệu được gắn nhãn
* Có thể tìm ra các mẫu dữ liệu mới và phát hiện các mối quan hệ mới giữa các đặc trưng
* Có thể áp dụng cho nhiều lĩnh vực khác nhau như phân cụm, liên kết, phân tích chuỗi thời gian và nhiều hơn nữa
* Nhược điểm:
* Không thể đưa ra dự đoán chính xác về tương lai.
* Không thể đưa ra kết quả chính xác khi không có đủ dữ liệu.
* Không thể kiểm soát được quá trình học và kết quả đầu ra
* Khó khăn trong việc đánh giá hiệu quả của thuật toán

***2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn***

- Mục đích của phương pháp: Phân cụm mức độ ưu tiên của bệnh nhân bằng thuật toán K-Means

- Input: Bộ dữ liệu đầu vào cho bài toán bao gồm các thuộc tính của các bệnh nhân, bao gồm: age, gender, chest pain type, blood pressure, cholesterol, max heart rate, exercise angina, plasma glucose, skin thickness, insulin, BMI, diabetes pedigree, hypertension, heart disease, residence type, smoking status.

- Output: Kết quả đầu ra là phân cụm của bệnh nhân dựa trên các yếu tố liên quan đến ưu tiên trong chăm sóc sức khỏe. Mỗi bệnh nhân được gán vào một cụm cụ thể, cho biết rằng họ thuộc vào nhóm có đặc điểm tương tự và ưu tiên tương tự.

- Method (Cách thực hiện):

* Đọc dữ liệu file.csv, chuẩn hóa dữ liệu
* Chia tập dữ liệu thành 2 phần: 90% dữ liệu sử dụng để huấn luyện mô hình, 10% dữ liệu còn lại được sử dụng để kiểm tra sự phù hợp của mô hình
* Chọn số cụm (K)
* Hiển thị kết quả và phân tích

- Các độ đo để đánh giá mô hình: Silhouette và Davies-Bouldin

**Phần 2. Thực nghiệm**

***1.Mô tả tập dữ liệu***

- Dữ liệu chứa thông tin của các bệnh nhân về các yếu tố liên quan đến sức khỏe bệnh nhân

Mỗi mẫu dữ liệu (vector) bao gồm các thuộc tính sau:

* age: Tuổi của bệnh nhân.
* gender: Giới tính của bệnh nhân (1 = nam, 0 = nữ).
* chest pain type: Loại đau ngực của bệnh nhân.
* blood pressure: Huyết áp tĩnh (mm Hg) khi nhập viện.
* cholesterol: Lượng cholesterol trong huyết thanh (mg/dl).
* max heart rate: Tần số tim tối đa đạt được.
* exercise angina: Tình trạng gây đau tim do tập thể dục (1 = có, 0 = không).
* plasma glucose: Mức đường huyết trong huyết thanh.
* skin thickness: Độ dày da.
* insulin: Lượng insulin.
* BMI: Chỉ số khối lượng cơ thể (Body Mass Index).
* diabetes pedigree: Hệ số dị diện đường huyết gia đình.
* hypertension: Tình trạng huyết áp cao (1 = có, 0 = không).
* heart disease: Bệnh tim (1 = có, 0 = không).
* residence type: Loại nơi cư trú (thành thị hoặc nông thôn).

Tập dữ liệu chứa thông tin từ 5109 bệnh nhân.

- Mô tả bài toán: Bài toán này nhằm mục đích phân loại hoặc phân cụm các bệnh nhân dựa trên các thông tin về các yếu tố ảnh hưởng đến mức độ ưu tiên của họ trong việc chăm sóc sức khỏe. Chúng ta sử dụng thuật toán K-Means để tạo ra các cụm (clusters) của bệnh nhân, với mỗi cụm bao gồm các bệnh nhân có đặc điểm tương tự về sức khỏe và các yếu tố liên quan đến ưu tiên trong chăm sóc sức khỏe. Mục tiêu của bài toán là hiểu cách các đặc điểm này có thể ảnh hưởng đến ưu tiên của bệnh nhân và xác định những nhóm ưu tiên khác nhau.

***2. Mô tả cách giải bài toán bằng phương pháp học máy***

Cách lựa chọn số cụm phù hợp nhất với tập dữ liệu của bài toán clustering là lựa chọn số cụm (K) phù hợp nhất với tập dữ liệu và đạt kết quả tốt nhất:

1. Chạy thuật toán K-Means với nhiều giá trị của K: Bắt đầu bằng việc chạy thuật toán K-Means với một loạt các giá trị K khác nhau (K từ 2 đến 10)

2. Đánh giá mỗi kết quả phân cụm: Sử dụng các độ đo (Silhouette Score và Davies-Bouldin Score) để đánh giá chất lượng của kết quả phân cụm cho mỗi giá trị K.

3. Chọn K tốt nhất: Dựa trên biểu đồ và độ đo, chọn giá trị K mà cho kết quả tốt nhất.(K mà cho Silhouette Score cao nhất và Davies-Bouldin Score thấp nhất là lựa chọn tốt nhất)

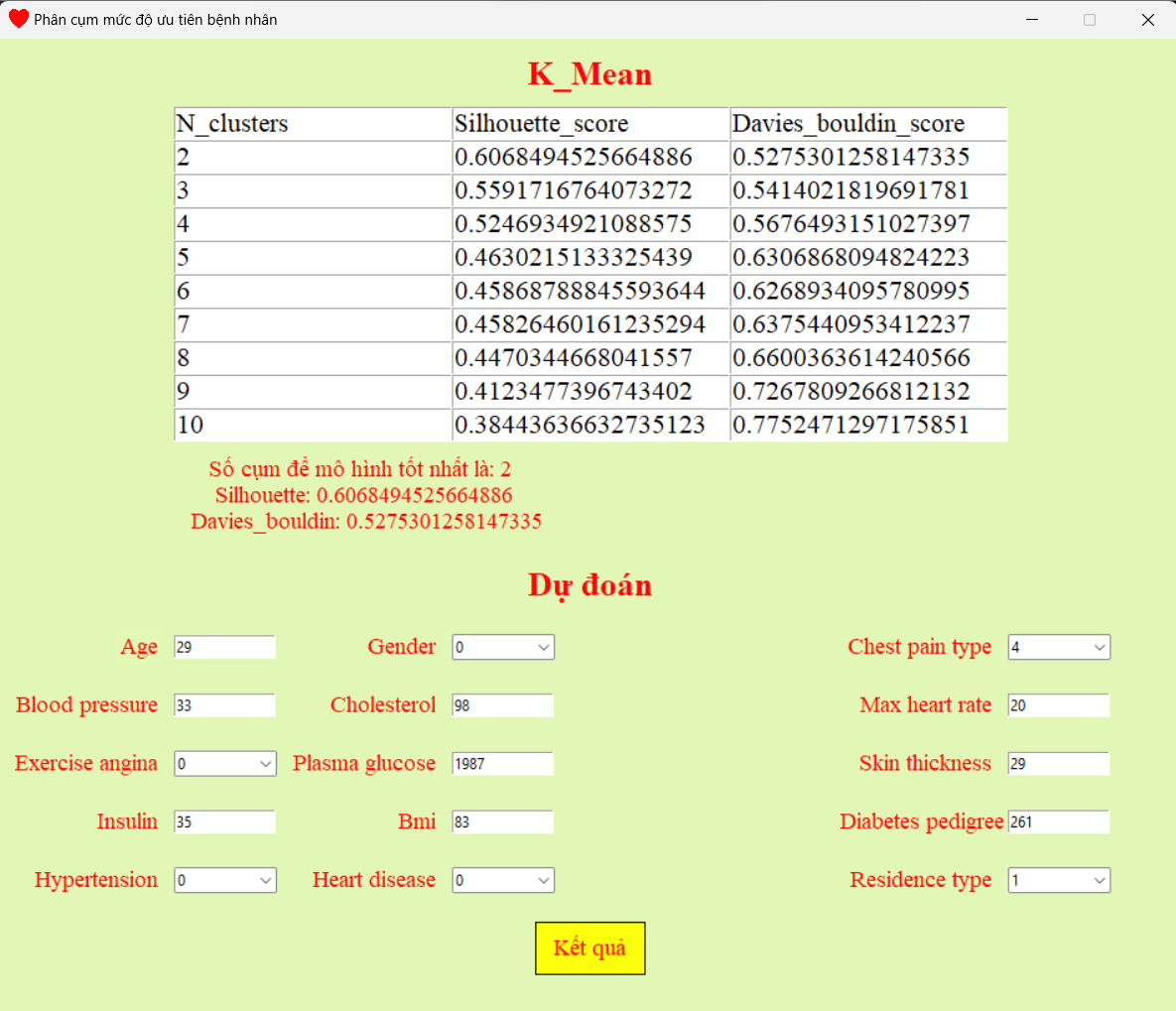
Lặp lại các bước trên để lựa chọn số cụm phù hợp nhất với tập dữ liệu của bài toán làm tối ưu hóa cả hai độ đo Silhouette và Davies-Bouldin để đảm bảo kết quả phân cụm tốt nhất cho đề tài.

***3. Đánh giá mô hình***

* Đánh giá chất lượng mô hình qua quan sát các độ đo Silhouette
  + Silhouette Score là một độ đo thường được sử dụng để đánh giá chất lượng của việc phân cụm trong bài toán K-Means. Nó đo lường độ tách biệt giữa các cụm và độ gần gũi của các điểm trong cùng một cụm.
  + Silhouette Score có giá trị nằm trong khoảng [-1, 1].
  + Giá trị lớn hơn 0 cho thấy rằng các điểm trong cùng một cụm gần nhau hơn so với các điểm trong các cụm khác. Giá trị gần 1 là tốt nhất, trong khi giá trị gần -1 chỉ ra rằng cụm không được phân tách tốt.
  + Khi đánh giá mô hình, cần tìm giá trị Silhouette Score cao nhất.
* Đánh giá chất lượng mô hình qua quan sát các độ đo Davies-Bouldin
  + Davies-Bouldin Score là một độ đo khác để đánh giá sự tách biệt giữa các cụm. Nó đo lường độ tách biệt giữa mỗi cụm và cụm gần nhất (cụm có độ tương tự nhau nhất).
  + Giá trị của Davies-Bouldin Score nằm trong khoảng [0, vô cùng).
  + Giá trị nhỏ hơn cho thấy rằng các cụm tách biệt tốt và có sự đồng nhất tốt hơn. Khi đánh giá mô hình, cần tìm giá trị Davies-Bouldin Score thấp nhất.
* Với bài toán đề ra ta thấy tại N\_clusters = 2 thì có độ đo Silhouette Score cao nhất và độ đo Davies-Bouldin Score thấp nhất.

=> Số cụm để cho mô hình của bài toán được tối ưu nhất là 2

***4. Mô tả các chức năng của chương trình***



* Chức năng của chương trình giúp cho có thể phân loại mức độ ưu tiên của các bệnh nhân với những căn bệnh, tình trạng sức khỏe khác nhau để có thể sắp xếp độ ưu tiên và cứu chữa kịp thời
* Giao diện chương trình gồm có:
* Age: Tuổi của bệnh nhân.
* Gender: Giới tính của bệnh nhân (1 = nam, 0 = nữ).
* Chest pain type: Loại đau ngực của bệnh nhân (Có các mức độ từ 0 đến 4)
* Blood pressure: Huyết áp tĩnh (mm Hg) khi nhập viện.
* Cholesterol: Lượng cholesterol trong huyết thanh (mg/dl).
* Max heart rate: Tần số tim tối đa đạt được.
* Exercise angina: Tình trạng gây đau tim do tập thể dục (1 = có, 0 = không).
* Plasma glucose: Mức đường huyết trong huyết thanh.
* Skin thickness: Độ dày da.
* Insulin: Lượng insulin.
* BMI: Chỉ số khối lượng cơ thể (Body Mass Index).
* Diabetes pedigree: Hệ số dị diện đường huyết gia đình.
* Hypertension: Tình trạng huyết áp cao (1 = có, 0 = không).
* Heart disease: Bệnh tim (1 = có, 0 = không).
* Residence type: Loại nơi cư trú (thành thị hoặc nông thôn).

**Phần 3. Kết luận**

* Qua bài toán phân cụm nhóm các bệnh nhân ưu tiên ứng dụng thuật toán phân cụm K-means thuộc loại học không giám sát. Nhóm em đã hiểu các bước thực hiện, từ việc tiền xử lý dữ liệu, chọn tham số, huấn luyện mô hình, sử dụng các độ đo và đưa ra kết luận và áp dụng được thuật toán vào bài toán.
* Trên đây là bản báo cáo của nhóm em nếu có gì sai sót mong cô có thể góp ý để nhóm em hoàn thiện hơn.

**Tài liệu tham khảo**

* Dataset trong bài toán: [Patient Priority for Clustering (kaggle.com)](https://www.kaggle.com/datasets/hossamahmedaly/patient-priority-classification)
* Thư viện sử dụng: [sklearn.cluster.KMeans — scikit-learn 1.3.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)
* Độ đo sử dụng:
* [sklearn.metrics.silhouette\_score — scikit-learn 1.3.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html)
* [sklearn.metrics.davies\_bouldin\_score — scikit-learn 1.3.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html)