BÁO CÁO KẾT QUẢ THỰC HÀNH LAB 3

I. Cơ sở lý thuyết

1) Tiền xử lý dữ liệu (EDA & Chuẩn hóa)

· Kiểm tra phân phối, giá trị thiếu/ngoại lai; mã hóa biến phân loại (One-Hot/Ordinal).

· Chuẩn hóa/Scale (StandardScaler, MinMaxScaler) khi mô hình nhạy khoảng cách (SVM, KNN) hoặc khi đặc trưng khác thang đo.

· Chia tập: train/valid/test; dùng StratifiedKFold cho dữ liệu mất cân bằng.

2) Chỉ số đánh giá

· Accuracy, Precision, Recall, F1; ROC-AUC, PR-AUC (hữu ích khi lệch lớp).

· Ma trận nhầm lẫn (confusion matrix) để phân tích lỗi.

· Đối với mô hình xác suất: hiệu chỉnh xác suất (Platt/Isotonic) nếu cần.

3) Decision Tree (DT)

· Ý tưởng: phân chia không gian bằng câu hỏi theo đặc trưng để tối đa hóa thông tin (Gini/Entropy).

· Ưu: dễ diễn giải, vẽ cây; Nhược: dễ overfit, nhạy nhiễu.

· Siêu tham số chính: max\_depth, min\_samples\_split, min\_samples\_leaf, max\_features, criterion (gini/entropy/log\_loss).

4) Random Forest (RF)

· Tập hợp nhiều cây (bagging) + chọn ngẫu nhiên đặc trưng → giảm phương sai, chống overfit.

· Có thể đo “feature importance” (Gini importance hoặc permutation importance).

· Siêu tham số: n\_estimators, max\_depth, max\_features, min\_samples\_leaf, bootstrap, class\_weight.

5) Support Vector Machine (SVM)

· Ý tưởng: biên cách ly tối đa; kernel trick cho biên phi tuyến (linear, rbf, poly, sigmoid).

· Bắt buộc chuẩn hóa dữ liệu; cân chỉnh C (độ phạt) và gamma (độ cong của RBF).

· Thích hợp khi số chiều lớn, số mẫu vừa phải.

6) Naive Bayes (NB)

· Giả định độc lập có điều kiện giữa các đặc trưng.

· GaussianNB: đặc trưng liên tục; MultinomialNB: đếm/tần suất (text); BernoulliNB: nhị phân.

· Nhanh, baseline tốt; nhược: giả định “ngây thơ” có thể không đúng.

7) Quy trình thực nghiệm (chuẩn)

· Pipeline(scaler, model) + GridSearchCV/StratifiedKFold.

· Lưu mô hình/báo cáo; vẽ ROC/PR, confusion matrix; so sánh mô hình theo nhiều chỉ số.

II. Phân công

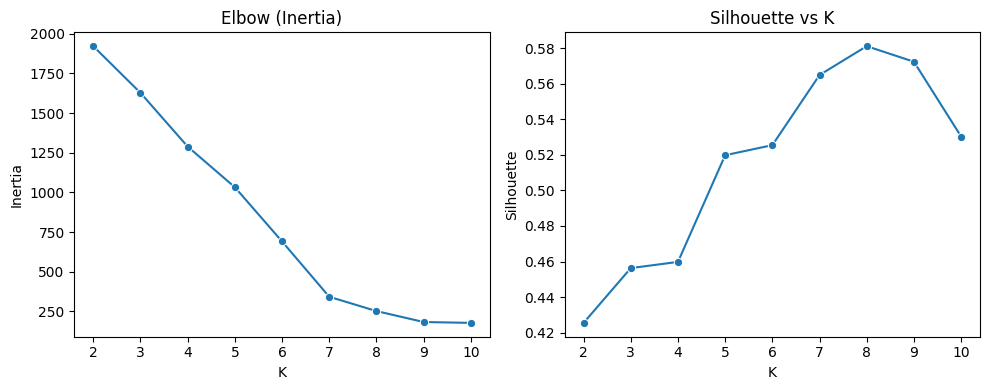
Bảng phân công (2 thành viên):

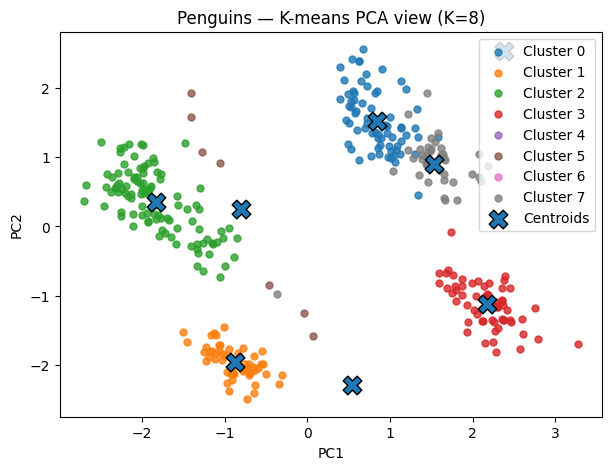
| Thành viên | Nhiệm vụ chính | Sản phẩm/Dầu ra |
| --- | --- | --- |
| Thành viên A | - Chuẩn bị & làm sạch dữ liệu (EDA, xử lý thiếu, mã hóa/scale).  - Xây dựng & tối ưu Decision Tree, Random Forest (GridSearchCV).  - Trực quan hóa: cây quyết định (Graphviz), feature importance RF. | - Notebook DT/RF hoàn chỉnh (Pipeline + CV).  - Biểu đồ ROC AUC, confusion matrix.  - Báo cáo ngắn DT/RF (1–2 trang). |
| Thành viên B | - Xây dựng & tối ưu SVM (chuẩn hóa bắt buộc, so sánh kernel).  - Xây dựng Naive Bayes (Gaussian/Multinomial/Bernoulli) theo loại dữ liệu.  - Trực quan hóa: ROC/PR, confusion matrix; so sánh mô hình. | - Notebook SVM + NB (Pipeline + CV).  - Bảng so sánh chỉ số (Accuracy, Precision, Recall, F1, ROC AUC).  - Báo cáo ngắn SVM/NB (1–2 trang). |

III. Bài tập thực hành

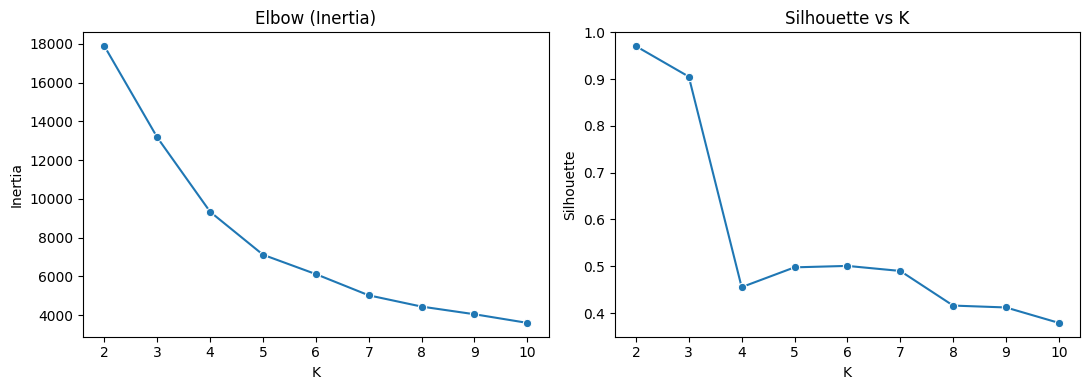
1.GIẢI THUẬT K-MEANS

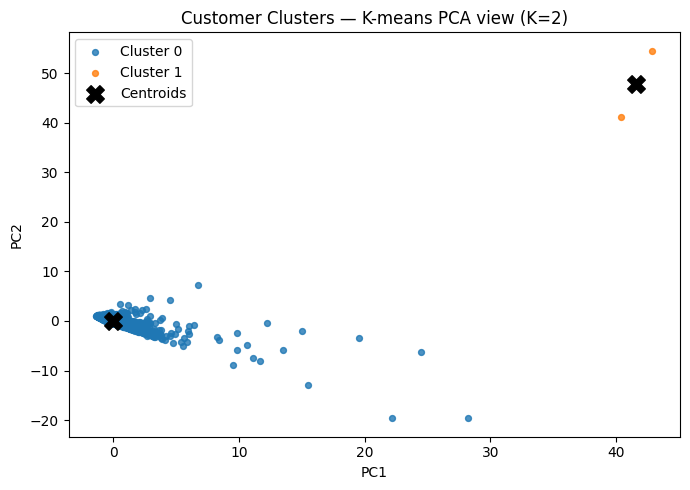
Phần A: dữ liệu chim cánh cụt





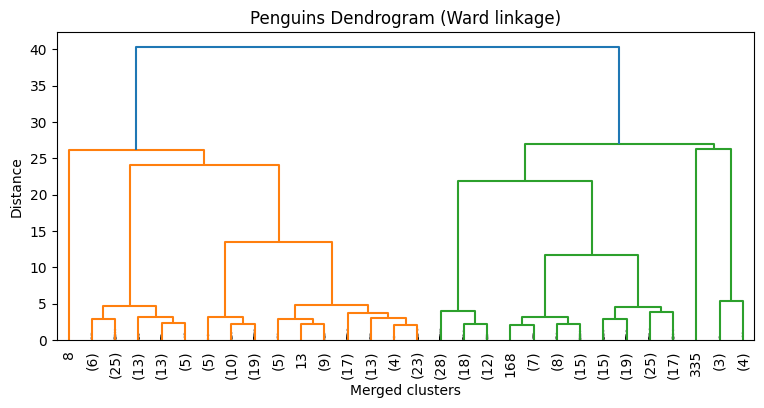
Phần B dữ liệu mua sắm tại siêu thị

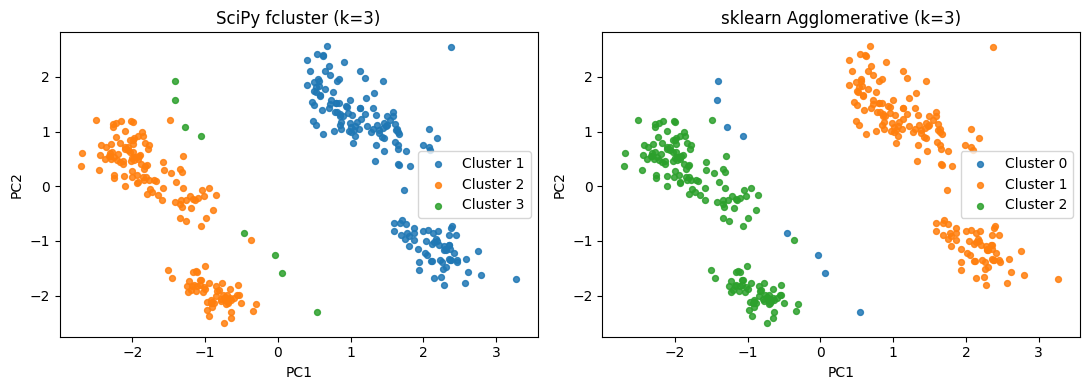




2.GIẢI THUẬT PHÂN CỤM ĐA CẤP

Phần A : dữ liệu chim cánh cụt





Phần B : dữ liệu mua sắm tại siêu thị

