

**ĐẠI HỌC HUẾ**

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

**🙢**

**BÁO CÁO ĐỒ ÁN**

**Học kỳ I, năm học 2023 - 2024**

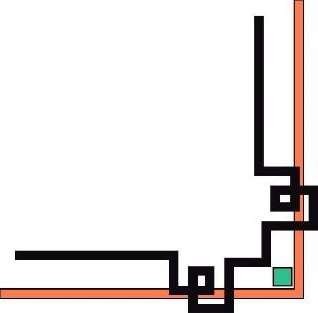
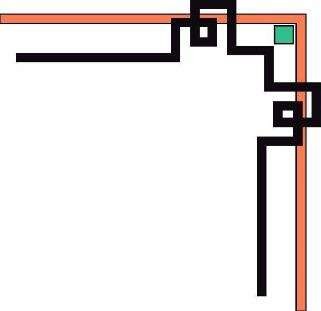
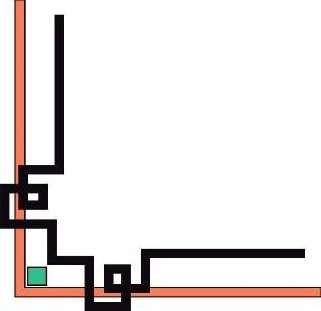
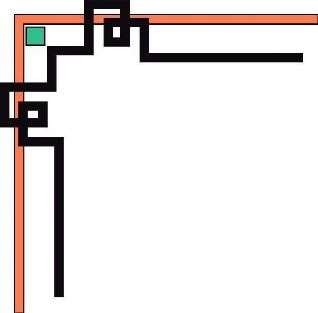
**Học phần:**

**HỌC MÁY 1**

**Số phách**

*(Do hội đồng chấm thi ghi)*

**Thừa Thiên Huế, tháng 01 năm 2024**



**ĐẠI HỌC HUẾ**

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

**🙢**

**BÁO CÁO ĐỒ ÁN**

**Học kỳ I, năm học 2023 – 2024**

**Học phần:**

**HỌC MÁY 1**

**Giảng viên hướng dẫn:** TS. Hoàng Hữu Trung

**Lớp:** Khoa học dữ liệu và trí tuệ nhân tạo - K3

**Sinh viên thực hiện:** Nguyễn Văn Minh Khánh

*(ký và ghi rõ họ tên)*

**Số phách**

*(Do hội đồng chấm thi ghi)*

**Thừa Thiên Huế, tháng 01 năm 2022**

# LỜI CẢM ƠN

“Em đã rất cố gắng và nỗ lực trong bài báo cáo đồ án này. Tuy nhiên, sẽ không thể thực hiện được nếu không có sự hỗ trợ, giúp đỡ ân cần của giảng viên bộ môn - Ngôn ngữ lập trình Python cũng như Ban giám hiệu Khoa Kỹ thuật và Công nghệ - Đại học Huế vì đã tạo điều kiện về cơ sở vật chất, như môi trường học tập thân thiện, giúp em phát huy hết khả năng học tập và rèn luyện nhân cách một cách hiệu quả.

Em muốn bày tỏ lòng biết ơn chân thành đối với giảng viên bộ môn - Ngôn ngữ lập trình Python và toàn thể giáo viên của Khoa Kỹ thuật và Công nghệ - Đại học Huế.

Em muốn bày tỏ lòng biết ơn đến gia đình và bạn bè vì đã luôn đồng hành, động viên và quan tâm em trên con đường học tập và trong cuộc sống.

Và lời cảm ơn đặc biệt cuối cùng em xin dành tặng cho bản thân chính mình vì đã không bỏ cuộc vào những lúc bản thân suy sụp, mệt mỏi nhất, cảm ơn bản thân đã luôn cố gắng để vượt qua những khó khăn tưởng chừng không thể bước tiếp, cảm ơn vì tất cả.”

# DANH MỤC HÌNH ẢNH

# DANH MỤC BẢNG BIỂU

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc155636822)

[DANH MỤC HÌNH ẢNH ii](#_Toc155636823)

[DANH MỤC BẢNG BIỂU iii](#_Toc155636824)

[MỤC LỤC iv](#_Toc155636825)

[CHƯƠNG 1: HỌC CÓ GIÁM SÁT (SUPERVISED LEARNING) 6](#_Toc155636826)

[1.1 Giới thiệu tập dữ liệu 6](#_Toc155636827)

[1.1.1 Tập dữ liệu 6](#_Toc155636828)

[1.1.2 Thành phần tập dữ liệu 6](#_Toc155636829)

[1.2 Đọc dữ liệu 7](#_Toc155636830)

[1.2.1 Nhập các thư viện cần thiết 7](#_Toc155636831)

[1.2.2 Nhập tập dữ liệu 8](#_Toc155636832)

[1.3 Đọc hiểu tập dữ liệu 8](#_Toc155636833)

[1.3.1 Thông tin dữ liệu 8](#_Toc155636834)

[1.3.2 Phân tách tập dữ liệu 10](#_Toc155636835)

[1.3.3 Kiểm tra giá trị ngoại lệ trong biến liên tục 15](#_Toc155636836)

[1.4 Tiền xử lý dữ liệu 16](#_Toc155636837)

[1.4.1 Kiểm tra và xử lý giá trị trùng lặp 16](#_Toc155636838)

[1.4.2 Xử lý giá trị bị thiếu 17](#_Toc155636839)

[1.4.3 Chuyển đổi dữ liệu 17](#_Toc155636840)

[1.4.4 Xử lý các giá trị ngoại lệ 18](#_Toc155636841)

[1.4.5 Lưu tập dữ liệu sau khi đã tiền xử lý 20](#_Toc155636842)

[1.5 Mối tương quan của tập dữ liệu 21](#_Toc155636843)

[1.5.1 Biểu đồ phân tán của cột age và thalachh 21](#_Toc155636844)

[1.5.2 Biểu đồ phân phối tần suất 21](#_Toc155636845)

[1.5.3 Biểu đồ tròn thể tỉ lệt của cột output so với cột sex 22](#_Toc155636846)

[1.5.4 Biểu đồ mật độ (KDE) của các trường dữ liệu so với output 23](#_Toc155636847)

[1.5.5 Biểu đồ thể hiện mức độ tương quan của toàn bộ tập dữ liệu 25](#_Toc155636848)

[1.6 Mô hình 26](#_Toc155636849)

[1.6.1 Chuẩn bị mô hình 26](#_Toc155636850)

[1.6.2 Linear Regression 27](#_Toc155636851)

[1.6.3 Logistic Regression 28](#_Toc155636852)

[1.6.4 Support Vector Machine 30](#_Toc155636853)

[1.6.5 K-Nearest Neighbours 33](#_Toc155636854)

[1.6.6 Bernoulli Naive Bayes 36](#_Toc155636855)

[1.7 cc 39](#_Toc155636856)

[CHƯƠNG 2: HỌC KHÔNG GIÁM SÁT (UNSUPERVISED LEARNING) 41](#_Toc155636857)

[TÀI LIỆU THAM KHẢO 43](#_Toc155636858)

[KẾT QUẢ KIỂM TRA ĐẠO VĂN 44](#_Toc155636859)

# CHƯƠNG 1: HỌC CÓ GIÁM SÁT (SUPERVISED LEARNING)

## 1.7 Phân tích, đánh giá và kết luận

### 1.7.1 Phân tích, đánh giá các thuật toán

|  |  |
| --- | --- |
|  | Linear Regression |
| MSE - Độ Lỗi Trung Bình Bình Phương | 0.1223 |
| R² - Hệ số xác định | 0.4903 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Logistic Regression | SVM | Knearest Neighbors | Bernoulli NB |
| Độ chính xác | 0.82 | 0.82 | 0.85 | 0.88 |

Linear Regression có MSE khá thấp (0.1223), đồng thời là 0.4903. Điều này có nghĩa là mô hình có khả năng giải thích được 49.03% biến thiên của dữ liệu, tức là nó tốt hơn so với một mô hình đoán trung bình.

Logistic Regression và SVM: cả hai đều có độ chính xác giống nhau, đạt 82%. Điều này cho thấy mô hình khá hiệu quả trong việc dự đoán nhóm.

KNN: có độ chính xác cao nhất, đạt 85% khi sử dụng k = 11.

Bernoulli Naive Bayes: có độ chính xác cao nhất, đạt 88.33%.

Phân tích và Đánh Giá Chung:

* Độ chính xác: Bernoulli Naive Bayes có độ chính xác tốt nhất trên tập dữ liệu.
* Thời gian xử lý: Linear Regression là nhanh nhất, nhưng Bernoulli Naive Bayes cũng có thời gian xử lý chấp nhận được.
* Phân tích mô hình: Linear Regression cho phép hiểu rõ tác động của từng biến độc lập lên biến phụ thuộc. Trong khi đó, các mô hình khác như Naive Bayes và k-NN cung cấp độ chính xác tốt mà không cần nhiều giả định về dữ liệu.

### 1.7.2 Kết luận

1. Độ Chính Xác: Bernoulli Naive Bayes là sự lựa chọn tốt nhất trong trường hợp này, với độ chính xác đạt 88.33%.
2. Các mối quan hệ giữa các trường dữ liệu: Nếu bạn muốn hiểu rõ hơn về mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc, Linear Regression có thể cung cấp thông tin chi tiết về tác động của từng biến.
3. Độ Phức Tạp Của Mô Hình: Nếu ưu tiên độ phức tạp thấp và tính dễ diễn giải, Linear Regression và Logistic Regression là lựa chọn tốt. K-NN và Naive Bayes đơn giản và dễ triển khai.
4. Thời Gian Xử Lý: Nếu tốc độ xử lý là một ưu tiên, Linear Regression và SVM là lựa chọn tối ưu. K-NN và Naive Bayes cũng có thời gian xử lý tốt, nhưng có thể tăng khi có nhiều dữ liệu.
5. Tập dữ liệu: tập dữ liệu này thuộc tập phân loại nên ưu tiên chọn thuật toán BernoulliNB, tiếp đến là k-NN, Logistic Regression, SVM và cuối cùng là Linear Regression.

.

# CHƯƠNG 2: HỌC KHÔNG GIÁM SÁT (UNSUPERVISED LEARNING)

## 1.7 Phân tích, đánh giá và kết luận

### 1.7.1 Phân tích, đánh giá các thuật toán

|  |  |  |  |
| --- | --- | --- | --- |
|  | Kmeans | Hierarchical Clustering | DBScan |
| Độ chính xác | 0.6044 | 0.2007 | 0.2677 |
| Thời gian (s) | 0.2576 | 0.007 | 0.002 |

1. Kmeans:
   1. Phân tích:
      1. Độ chính xác tương đối nhưng không cao.
      2. Thời gian chạy ổn định nhưng không phải là nhanh nhất.
   2. Đánh giá:
      1. K-Means là một phương pháp tốt cho việc phân cụm, nhưng có thể cần điều chỉnh số lượng cụm để cải thiện độ chính xác.
2. Hierarchical Clustering:
   1. Phân tích:
      1. Độ chính xác thấp, có thể không phù hợp cho dữ liệu có cấu trúc phức tạp.
      2. Thời gian chạy nhanh.
   2. Đánh giá:
      1. Hierarchical Clustering không đạt được độ chính xác cao, có thể không phù hợp cho các tập dữ liệu phức tạp.
3. DBSCAN:
   1. Phân tích:
      1. Độ chính xác tương đối cao, đặc biệt là so với Hierarchical Clustering.
      2. Thời gian chạy rất nhanh.
   2. Đánh giá:
      1. DBSCAN có thể là lựa chọn tốt với dữ liệu có cấu trúc phức tạp, với độ chính xác khá và thời gian chạy ngắn.

### 1.7.2 Kết luận

* + - Nếu ưu tiên độ chính xác:
      * + Chọn DBSCAN, với độ chính xác tương đối cao.
    - Nếu ưu tiên thời gian chạy:
      * + Chọn DBSCAN, với thời gian chạy ngắn nhất.
    - Nếu cần cả hai:
      * + DBSCAN vẫn là lựa chọn tốt nhất với sự cân bằc giữa độ chính xác và thời gian chạy.

=> Đối với tập dữ liệu này nên chọn thuật toán DBSCAN vì

* DBSCAN có thời gian chạy rất nhanh, làm cho nó là lựa chọn hữu ích cho việc xử lý dữ liệu lớn.
* DBSCAN thường hiệu quả với dữ liệu có các cụm có hình dạng và kích thước khác nhau.
* DBSCAN là một thuật toán linh hoạt và thời gian chạy nhanh, đặc biệt là khi sử dụng cùng với Nearest Neighbors để tính khoảng cách
* DBSCAN là một lựa chọn mạnh mẽ đối với dữ liệu có cấu trúc phức tạp và không đồng đều.

# TÀI LIỆU THAM KHẢO

# KẾT QUẢ KIỂM TRA ĐẠO VĂN

Text

Description automatically generated with low confidence

**ĐẠI HỌC HUẾ**

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙢

**PHIẾU ĐÁNH GIÁ ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN**

**Học kỳ I, năm học 2021 – 2022**

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1** | **Cán bộ chấm thi 2** |
| **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  Điểm đánh giá của CBCT1:  Bằng số: ..............................................  Bằng chữ: ............................................ | **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  Điểm đánh giá của CBCT2:  Bằng số: ..............................................  Bằng chữ: ............................................ |

Điểm kết luận: ..................................................................................................................

Bằng số: ............................................................................................................................

Bằng chữ: ..........................................................................................................................

*Thừa Thiên Huế, ngày 15 tháng 01 năm 2022*

**Cán bộ chấm thi 1**

*(Ký và ghi rõ họ tên)*

**Cán bộ chấm thi 2**

*(Ký và ghi rõ họ tên)*