BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**🙞 🕮 🙜**



BỘ MÔN KHOA HỌC MÁY TÍNH

**Học phần KHAI KHOÁNG DỮ LIỆU**

**Đề tài**

**PHÂN TÍCH TẬP DỮ LIỆU ICE-CREAM**

Giáo viên hướng dẫn: Sinh viên thực hiện:

TS. Lưu Tiến Đạo 1. Nguyễn Việt Hào B1812338

2. Nguyễn Phúc Nguyên B1812364

Cần Thơ, 11/2021

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**🙞 🕮 🙜**



BỘ MÔN KHOA HỌC MÁY TÍNH

**Học phần KHAI KHOÁNG DỮ LIỆU**

**Đề tài**

**PHÂN TÍCH TẬP DỮ LIỆU ICE-CREAM**

Giáo viên hướng dẫn: Sinh viên thực hiện:

TS. Lưu Tiến Đạo 1. Nguyễn Việt Hào B1812338

2. Nguyễn Phúc Nguyên B1812364

Cần Thơ, 11/2021

**NHẬN XÉT CỦA GIẢNG VIÊN**

**--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**MỤC LỤC**

**DANH MỤC HÌNH**

[Hình 1 Ảnh minh họa - ảnh gương dọc 8](#_Toc497570970)

[Hình 2 Ảnh kết quả - ảnh gương dọc 9](#_Toc497570971)

[Hình 3 Ảnh minh họa - ảnh gương ngang 9](#_Toc497570972)

[Hình 4 Ảnh kết quả - ảnh gương ngang 10](#_Toc497570973)

[Hình 5 Ảnh minh họa - ảnh viết chì 10](#_Toc497570974)

[Hình 6 Ảnh kết quả - viết chì thường 11](#_Toc497570975)

[Hình 7 Ảnh kết quả - viết chì thường 11](#_Toc497570976)

[Hình 8 Ảnh minh họa – tranh sơn dầu 12](#_Toc497570977)

[Hình 9 Ảnh kết quả - tranh sơn dầu 13](#_Toc497570978)

[Hình 10 Ảnh minh họa – thêm chữ 13](#_Toc497570979)

[Hình 11 Ảnh minh họa - ảnh cầu vồng 14](#_Toc497570980)

[Hình 12 Ảnh minh họa – Cool effect 14](#_Toc497570981)

[Hình 13 Giao diện sử dụng 24](#_Toc497570982)

[Hình 14 Giao diện chụp ảnh 24](#_Toc497570983)

**DANH MỤC BẢNG**

[Bảng 1 Ma trận pixel gốc 15](#_Toc497534246)

[Bảng 2 Ma trận pixel kết quả – gương ngang 15](#_Toc497534247)

[Bảng 3 Ma trận pixel kết quả - gương dọc 15](#_Toc497534248)

## Giới thiệu dữ liệu

Dữ liệu ice-cream được thu thập trên 200 học sinh trung học và là điểm số của các bài kiểm tra khác nhau, bao gồm một trò chơi điện tử và một câu đố. Kết quả đo lường trong phân tích này là hương vị yêu thích của kem - vani, sô cô la hoặc dâu tây- từ đó chúng ta sẽ xem mối quan hệ nào tồn tại với điểm trò chơi điện tử (video), điểm câu đố (câu đố) và giới tính (nữ).

## Tìm hiểu dữ liệu

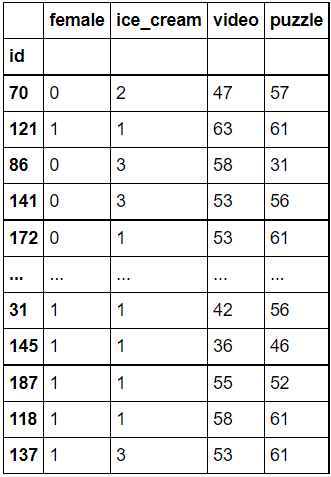
|  |  |  |
| --- | --- | --- |
| Variable name | Variable | Data type |
| Id | Định danh học sinh | Nominal |
| female | Gới tính(0: Male, 1: Female) | Binary |
| ice\_cream | Hương vị yêu thích (1: Vanilla, 2: Chocolate, 3: Strawberry) | Nominal |
| video | Điểm trò chơi điện tử | Scale/Continuous |
| puzzle | Điểm câu đố | Scale/Continuous |

Bảng: Mô tả tập dữ liệu ice-cream

Tập dữ liệu bao gồm 200 phần tử, 3 cột thuộc tính, 1 cột nhãn

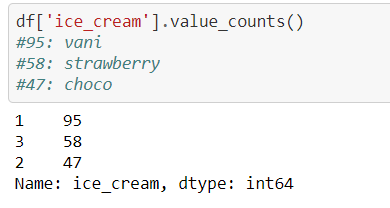
* female, video, puzzle là các thuộc tính
* ice\_cream là nhãn

Vì nhãn nhận các giá trị rời rạc 1, 2, 3 nên đây là bài toán phân lớp.

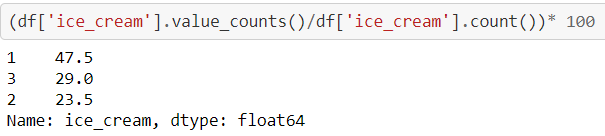


Bảng: Tập dữ liệu ice-cream

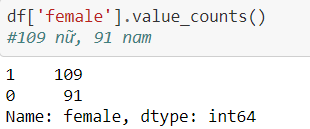
Vani là nhóm kem phổ biến nhất:



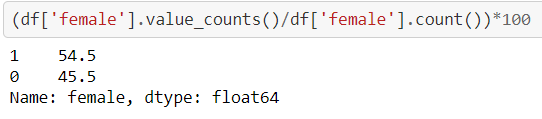
Phần trăm của các nhãn:



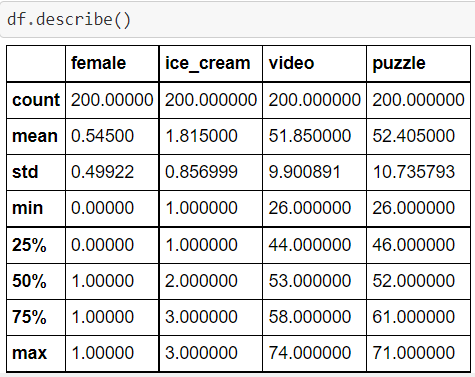
Số lượng nam và nữ:



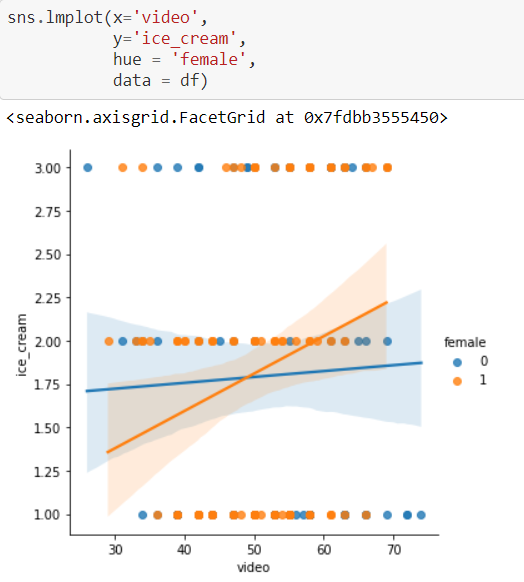
Tỉ lệ nam và nữ:



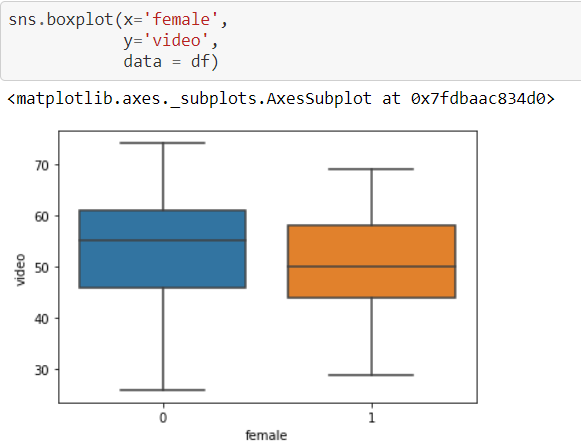
Mô tả dữ liệu:



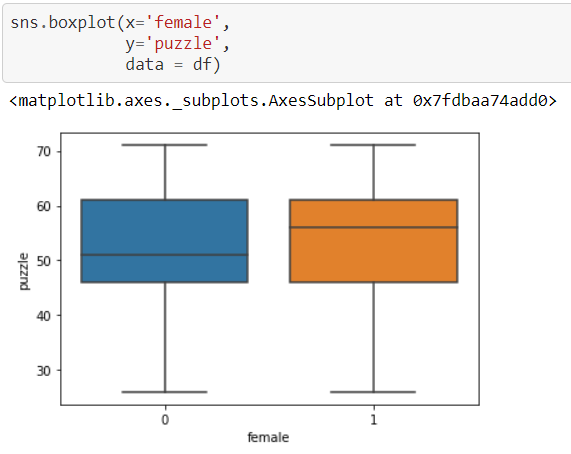
## Phân tích mối quan hệ giữa các biến



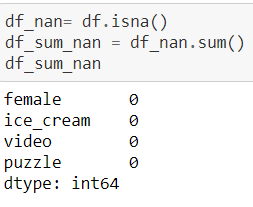
Quan hệ giữa female và video:



Quan hệ giữa female và puzzle:



Dữ liệu sai lệch:



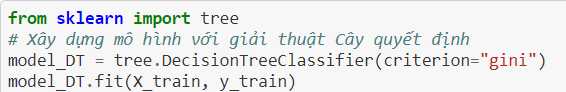
## Xây dựng mô hình

Sử dụng nghi thức kiểm tra Hold-out chia dữ liệu ngẫu nhiên thành 2 tập dữ liệu con training set và test set theo tỷ lệ 70/30

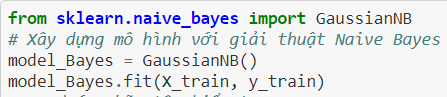


Sử dụng thư viện sklearn để xây dựng các mô hình phân lớp

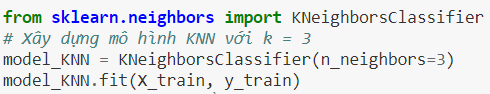
* Xây dựng mô hình với giải thuật cây quyết định:



* Xây dựng mô hình với giải thuật Naive Bayes:



* Xây dựng mô hình với giải thuật KNN:



## Đánh giá kết quả

Giải thuật Naive Bayes cho độ chính xác 53.3%, là giải thuật cho độ chính xác cao nhất. Giải thuật cây quyết định và KNN cho độ chính xác là 50%.

Bảng: So sánh độ chính xác các giải thuật