N20DCCN020-VŨ HUY HÙNG

N20DCCN022-VÕ QUANG HUY

N20DCCN026-VĂN TỐ HỮU

**Báo cáo chuyên đề 6**

1. Tự cài đặt 2 thuật toán trên.

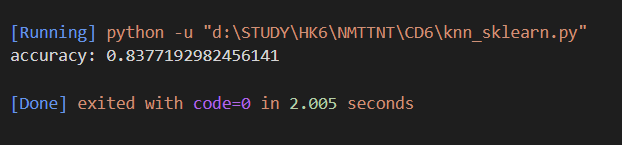
2. Sử dụng phần cài đặt thuật toán của sklearn (hoặc thư viện khác), so sánh kết quả của KNN và Decision Trees.

So sánh 1:

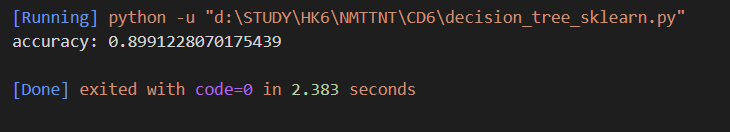
Testset: 228 mẫu

Trainset: 1500 mẫu

KNN



Decision tree



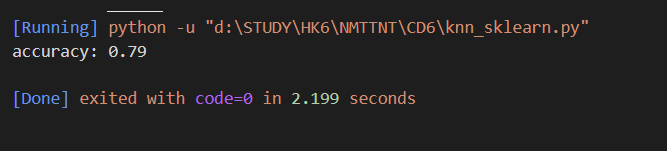
Đánh giá: Độ chính xác Decision Tree lớn hơn KNN và Decision Tree có tốc độ chạy nhanh hơn

So sánh 2:

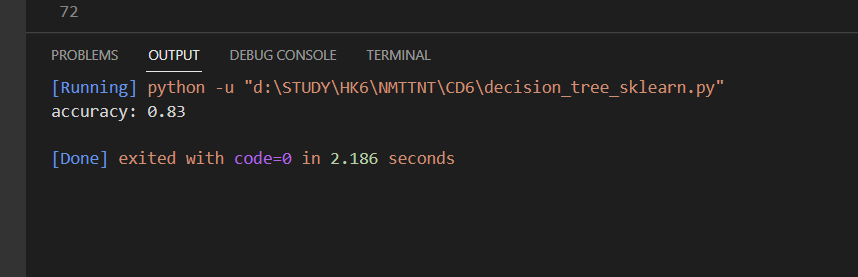
Testset: 100 mẫu

Trainset: Giữ nguyên

KNN



Decision Tree



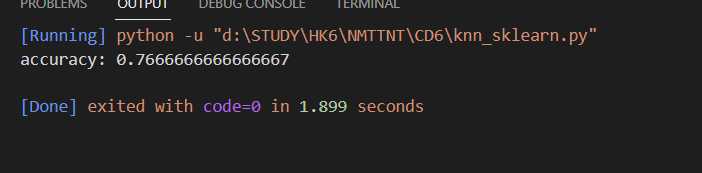
Đánh giá: Độ chính xác Decision Tree lớn hơn KNN và có tốc độ gần như bằng nhau

So sánh 3:

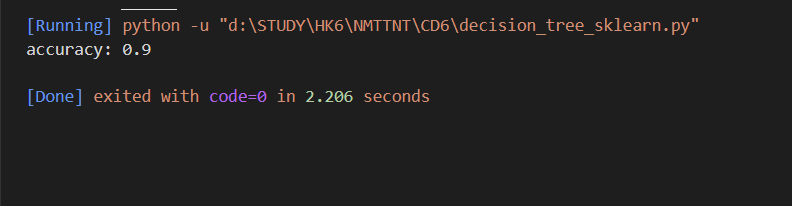
Testset: 30 mẫu

Trainset: Giữ nguyên

KNN



Decision Tree



Đánh giá: Độ chính xác Decision Tree lớn hơn KNN và KNN có tốc độ lớn hơn.

Giải thích: Vì KNN không hẳn là tạo mô hình train, nó chỉ so sánh từng mẫu trong testset với tất cả mẫu trong trainset, nên dữ liệu tet càng nhiều thì KNN chạy càng lâu.

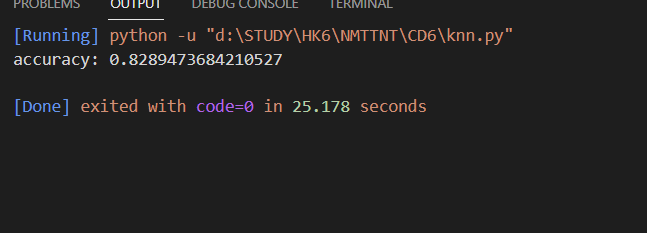
3. So sánh kết quả giữa thuật toán sinh viên tự cài đặt và thuật toán của thư viện, giải thích nếu có sự khác nhau.

Testset: 228 mẫu.

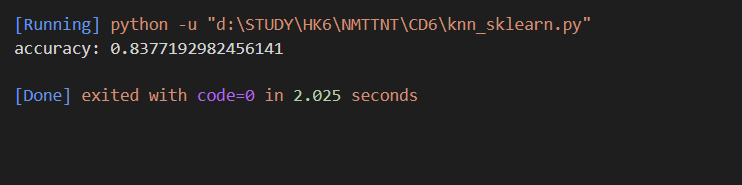
Trainset: 1500 mẫu.

So sánh KNN tự cài đặt (thước đo euclid) và sử dụng thư viện sklearn , chọn 10 mẫu láng giềng gần nhất.

Tự cài đặt:



Dùng Sklearn



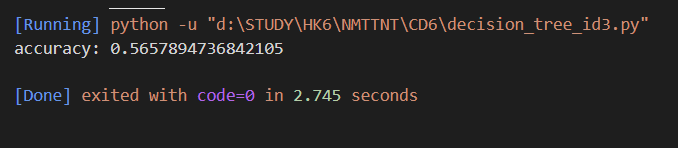
Đánh giá: Dùng thư viện Sklearn có tốc độ nhanh hơn nhiều và độ chính xác cũng cao hơn

Giải thích: Sklearn dùng công thức minkowski để tính khoảng cách thay vì euclid (đoán: có thể đo khoảng cách trong không gian nhiều chiều tốt hơn)

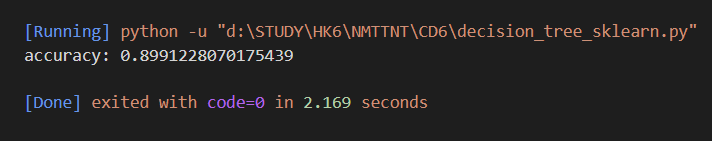
Sklearn dùng 3 thuật toán ball tree, kd tree, brute. Nó sẽ chuyển linh động thuật toán dựa vào đầu vào đưa ra để tính k láng giềng gần nhất.

So sánh Decision Tree tự cài đặt bằng ID3 và sử dụng thư viện sklearn

Tự cài đặt:



Dùng Sklearn



Đánh giá: Dùng thư viện Sklearn độ chính xác cao hơn và tốc độ cũng nhanh hơn.

Giải thích:

Sklearn default dùng Gini thay vì Entropy để tính Information Gain

Có thể Sklearn sử dụng một vài thuật toán tối ưu hơn mà không phải ID3

4. So sánh kết quả của KNN và Decision Trees (lưu ý, test set cần giống nhau để kết quả so sánh chính xác), nhận xét ưu và nhược điểm của 2 thuật toán, khi nào nên dùng KNN, khi nào nên dùng Decision Trees.

### Ưu điểm của KNN

Độ phức tạp tính toán của quá trình training là bằng 0.

Việc dự đoán kết quả của dữ liệu mới rất đơn giản.

Không cần giả sử gì về phân phối của các class.

### Nhược điểm của KNN

KNN rất nhạy cảm với nhiễu khi K nhỏ.

Như đã nói, KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó việc tính khoảng cách tới *từng* điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.

### Ưu điểm Decision tree

Mô hình dễ hiểu và dễ giải thích.

Cần ít dữ liệu để huẩn luyện.

Có thể xử lý tốt với dữ liệu dạng số (rời rạc và liên tục) và dữ liệu hạng mục.

Mô hình dạng white box rõ ràng.

Xây dựng nhanh.

Phân lớp nhanh.

### Nhược điểm Decision tree

Không đảm bảo xây dựng được cây tối ưu.

Có thể overfitting (tạo ra những cây quá khớp với dữ liệu huấn luyện hay quá phức tạp).

Thường ưu tiên thuộc tính có nhiều giá trị (khắc phục bằng các sử dụng Gain Ratio).