BÀI TẬP KIỂM TRA BUỔI 3 – CÂY QUYẾT ĐỊNH
Chiều 2

1. Xây dựng cây quyết định dựa vào **chỉ số độ lợi thông tin** và dự đoán nhãn

   a. Đọc dữ liệu từ tập dữ liệu đánh giá chất lượng rượu vang **trắng** trên trang UCI https://archive.ics.uci.edu/ml/datasets/wine+quality

   b. . Dữ liệu có bao nhiêu thuộc tính? Cột nào là cột nhãn? Giá trị của các nhãn (ghi lại kết quả và code vào file nộp bài)

   c. Với tập dữ liệu wineWhite sử dụng nghi thức K-Fold để phân chia tập dữ liệu huấn luyện với K=50, sử dụng tham số "Shuffle" để xáo trộn tập dữ liệu trước khi phân chia. Xác định số lượng phần tử có trong tập test và tập huấn luyện nếu sử dụng nghi thức đánh giá này.

   d. Xây dựng mô hình cây quyết định dựa trên tập dữ liệu học tạo ra ở bước c.

   e. Đánh giá độ chính xác cho từng phân lớp dựa vào giá trị dự đoán của câu 3 cho mỗi lần lặp. Chép lại kết quả độ chính xác cho từng phân lớp của lần lặp cuối nộp lại (có thể đưa vào comment trong file code).

   f. Tính độ chính xác tổng thể cho mỗi lần lặp và độ chính xác tổng thể của trung bình **50 lần lặp**

   g. Sử dụng giải thuật KNN, Bayes thơ ngây ở buổi thực hành số 2 để so sánh hiệu quả phân lớp của giải thuật cây quyết định với nghi thức đánh giá **k-fold với K=60**

2. Cho tập dữ liệu gồm 5 phần tử như bảng bên dưới,

STT	Chiều cao	Độ dài mái tóc	Giọng nói	**Nhãn**
1.	180	15	0	**0**
2.	167	42	1	**1**
3.	136	35	1	**1**
4.	174	15	0	**0**
5.	141	28	1	**1**

   - Sử dụng 5 phần tử trong tập dữ liệu trên để xây dựng mô hình dựa vào **chỉ số độ lợi thông tin** với thuộc tính: chiều cao, độ dài tóc và giọng nói để dự đoán nhãn là nam giới hay nữa giới.
   - Dự báo phần tử mới tới có thông tin chiều cao=135, độ dài mái tóc = 39 và giọng nói có giá trị là 1 thì người này là 0 hay 1?