# Bài tập thực hành:
# Phân Lớp (Classification)

*Bộ môn: KHMT*

Thời gian làm bài: 2 tuần (Xem dealine trong link nộp bài trên moodle)

Nộp bài:

- Nộp bài lên moodle.
- Đặt tên bài nộp theo định dạng MSSV1_MSSV2.rar. Trong đó bao gồm:
  - Tập tin báo cáo.
  - Các tập tin dữ liệu theo yêu cầu của bài tập

*Các hành vi sử dụng toàn bộ/một phần bài làm của người khác sẽ bị 0 điểm cho toàn bộ phần thực hành*

## Bài 1. Trả lời ngắn gọn các câu hỏi sau:

a. Tại sao phân lớp Bayes (Bayesian classification) được gọi là "naive" ("ngây thơ")?
b. Tại sao cần có bước tỉa nhánh (tree pruning) trong cây quyết định?
c. Khi dùng cây quyết định thì có thể sẽ bị overfitting. Overfitting là gì? Và nêu một vài phương pháp để hạn chế overfitting khi dùng cây quyết định.

## Bài 2. Phân Lớp.

Cho dữ liệu Mushroom (các loại nấm – xem link trên moodle). Thuộc tính cần phân lớp là nấm độc (p) hay không độc (ăn được - e).
**Mô tả:** có một vài giá trị không có trong file dữ liệu vì dữ liệu lớn và đã được chọn random thành tập nhỏ như bài tập.

| Tên thuộc tính | Mô tả | Giá trị |
|---|---|---|
| Class | Lớp phân loại | e = ăn được<br>p = có độc |
| Shape | Hình dáng | b = hình chuông,<br>c = hình nón<br>x = lồi<br>f = phẳng<br>k = có u<br>s = lõm |
| Suface | Bề mặt | f = có xơ<br>g = có đường rãnh (nếp)<br>y = có vảy<br>s = trơn |
| Color | Màu sắc | n = màu nâu |

| | | |
|---|---|---|
| | | b = màu da bò<br>c = màu nâu vàng<br>r = màu xanh lá<br>p = màu hồng<br>u = màu tím<br>e = màu đỏ<br>w = màu trắng<br>y = màu vàng |
| Bruise | Có vết thâm không | t = bị thâm<br>f = không màu |
| Odor | Có mùi hương | a = mùi quả hạnh<br>l = mùi cây hồi<br>c = mùi crê-ô-dốt<br>y = mùi cá<br>f = mùi thối<br>m = mùi mốc<br>n = không mùi<br>p =mùi hăng<br>s = mùi gia vị |
| Spore_Color | Màu bào tử (nầm) | k = màu đen<br>n = màu nâu<br>b = màu da bò<br>h = màu sô-cô-la<br>r = màu xanh lá<br>o = màu cam<br>u = màu tím<br>w = màu trắng<br>y = màu vàng |
| Population | Sự phân bố | s = mọc phân tán, rải rác<br>c = mọc chùm, cụm<br>n = mọc đông đảo, dày đặc<br>a = mọc phong phú<br>v = mọc nhiều<br>y = mọc đơn độc |
| Habitat | Môi trường sinh sống | g = đồng cỏ<br>l = lá cây<br>m = đồng cỏ (và các loại cây không phải thân gỗ khác)<br>p = đường đi<br>u = đô thị<br>w = bãi rác<br>d = thân gỗ |

1. Dùng weka chay Naïve Bayes cho tập train.csv để phân lớp một loại nấm là không có độc (e) hay có độc (p).
2. Từ model của tập train đã chạy được, chạy weka cho tập test.csv. So sanh kết quả của lớp phân lớp với kết quả tập test.csv và tính xem độ chính xác của phân lớp Naïve Bayes.

## Bài 3. SVM

Thực hiện các yêu cầu bài 2 với thuật toán phân lớp SVM.

## Bài 4: kNN

Cài đặt thuật toán kNN với các yêu cầu sau:

- **Input:**
  - ✓ *U* là mẫu cần phân lớp.
  - ✓ *T* là tập huấn luyện: $T = (t_{1\_1}, t_{1\_2}, t_{1\_3}, \dots, t_{1\_n}), \dots, (t_{m\_1}, t_{m\_2}, t_{m\_3}, \dots, t_{m\_n})$
  - ✓ Thuộc tính $t_{i\_n}$ là nhãn (label) của $T_i$
  - ✓ *m* là số lượng mẫu trong tập huấn luyện
  - ✓ *n* là số lượng thuộc tính trong mỗi mẫu.
  - ✓ *k* là số lượng láng giềng gần nhất ta cần tìm
- **Output**: Lớp của mẫu U

## Bài 5: (Cộng Điểm)

1. Dùng C/C++/C# để thực hiện Naïve Bayes. Rồi so sánh kết quả đạt được và kết quả chạy từ Weka giống nhau hay khác nhau? Vì sao?
2. Thay vì Naïve Bayes, hãy dùng kNN ở bài 4 để chạy cho dữ liệu Mushroom.

## Qui định:

- Làm bài theo nhóm. Mỗi nhóm tối đa 2 sinh viên.
- Hạn nộp: xem trên Moodle
- Bài nộp gồm file pdf/doc/docx trả lời câu hỏi lý thuyết, có đánh giá công việc từng cá nhân trong nhóm + thư mục source code.
- Đặt tất cả các nội dung được yêu cầu nộp trong thư mục có tên MSSV1_MSSV2, nén lại thành tập tin .zip hoặc .rar. Đại diện thay mặt nhóm để nộp ở link tương ứng trên Moodle.
- Sinh viên có thể viết chương trình bằng ngôn ngữ C/C++/C#.
- Các bài làm giống nhau hay chép code từ nơi khác sẽ bị 0 điểm.