**BÀI TẬP VỀ NHÀ – CHỦ ĐỀ: CÂY QUYẾT ĐỊNH**

**Thực hiện: Nhóm 14**

Bài 1: Dùng entropy để vẽ cây quyết định với dataset cho trong file hw\_dataset\_1.csv, biến mục tiêu là HEART ATTACK. Lưu ý PATIENT ID là thuộc tính dạng rời rạc.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PATIENT ID** | **CHEST PAIN** | **GENDER** | **SMOKES** | **EXERCISES** | **HEART ATTACK** |
| 1 | Yes | Male | No | Yes | Yes |
| 2 | Yes | Male | Yes | No | Yes |
| 3 | No | Female | Yes | No | Yes |
| 4 | No | Male | No | Yes | No |
| 5 | Yes | Female | Yes | Yes | Yes |
| 6 | No | Male | Yes | Yes | No |
| 7 | No | Female | No | Yes | No |

1. **Bước 1:** Tính Entropy cho biến mục tiêu “HEART ATTACK”
2. **Bước 2:** Tính Entropy trên các biến phụ thuộc
   1. **Biến “CHEST PAIN”:**

Với biến CHEST PAIN:

+ Có 3 dòng nhận giá trị “YES”, trong đó: 3 dòng “YES” ở biến mục tiêu

+ Có 4 dòng nhận giá trị “NO”, trong đó: 3 dòng “NO” ở biến mục tiêu và 1 dòng “YES” ở biến mục tiêu

* 1. **Biến “GENDER”:**

Với biến GENDER:

+ Có 3 dòng nhận giá trị “FEMALE”, trong đó: 2 dòng “YES” ở biến mục tiêu, 1 dòng “NO” ở biến mục tiêu

+ Có 4 dòng nhận giá trị “MALE”, trong đó: 2 dòng “NO” ở biến mục tiêu và 2 dòng “YES” ở biến mục tiêu

* 1. **Biến “SMOKES”:**

Với biến SMOKES:

+ Có 3 dòng nhận giá trị “NO”, trong đó: 1 dòng “YES” ở biến mục tiêu, 2 dòng “NO” ở biến mục tiêu

+ Có 4 dòng nhận giá trị “YES”, trong đó: 3 dòng “YES” ở biến mục tiêu và 1 dòng “NO” ở biến mục tiêu

* 1. **Biến “EXCERCISE”:**

Với biến EXCERCISE:

+ Có 2 dòng nhận giá trị “NO”, trong đó: 2 dòng “YES” ở biến mục tiêu

+ Có 5 dòng nhận giá trị “YES”, trong đó: 2 dòng “YES” ở biến mục tiêu và 3 dòng “NO” ở biến mục tiêu

* 1. **So sánh chỉ số G(S, thuộc tính):**

Để lấy độ lợi thông tin lớn nhất. Theo cách chọn “Entropy”, thuộc tính có độ lợi thông tin G(S, thuộc tính) lớn nhất thì ta chọn thuộc tính đó là thuộc tính phân lớp.

Ta có: *, , ,*

Có thể thấy lớn nhât nên chọn CHEST PAIN là thuộc tính phân lớp ở node đầu tiên.

**Tổng quan Cây quyết định sau bước 2:** A black background with blue orange and green squares

Description automatically generated

1. **Bước 3: Đi vào từng nhánh của CHEST PAIN**
   1. **Xét trên nhánh CHEST PAIN = “no”** 
      1. **Tính entropy với thuộc tính CHEST PAIN**

* Số lượng dữ liệu S = 4
* Số lượng HEART ATTACK = “yes”: 1
* Số lượng HEART ATTACK = “no”: 3
* Tính entropy:
  + 1. Biến “GENDER”:

Với biến GENDER:

+ Có 2 dòng nhận giá trị “FEMALE”, trong đó: 1 dòng “YES” ở biến mục tiêu, 1 dòng “NO” ở biến mục tiêu

+ Có 2 dòng nhận giá trị “MALE”, trong đó: 2 dòng “NO” ở biến mục tiêu.

* + 1. Biến “SMOKES”:

Với biến SMOKES:

+ Có 2 dòng nhận giá trị “NO”, trong đó: 2 dòng “NO” ở biến mục tiêu

+ Có 2 dòng nhận giá trị “YES”, trong đó: 1 dòng “YES” ở biến mục tiêu và 1 dòng “NO” ở biến mục tiêu

* + 1. Biến “EXERCISES”:

Với biến EXCERCISE:

+ Có 1 dòng nhận giá trị “NO”, trong đó: 1 dòng “YES” ở biến mục tiêu

+ Có 3 dòng nhận giá trị “YES”, trong đó: 3 dòng “NO” ở biến mục tiêu

* + 1. So sánh

Để lấy độ lợi thông tin lớn nhất. Theo cách chọn “Entropy”, thuộc tính có độ lợi thông tin G(S, thuộc tính) lớn nhất thì ta chọn thuộc tính đó là thuộc tính phân lớp.

Ta có: *, ,*

Có thể thấy lớn nhât nên chọn EXERCISE là thuộc tính phân lớp ở node tiếp theo ở nhánh CHEST PAIN = “no”.

Tổng quan cây quyết định lúc này:

A group of rectangular colored boxes with text

Description automatically generated

* 1. **Xét trên nhánh CHEST PAIN = “yes”**

Do chỉ ở nhánh này, target là class = “yes”. Nên không cần nhánh phân nhánh tiếp tục ở nhánh này

Nhận xét: Đến đây, các dữ liệu đã được phân chia đầy đủ vào các node lá. Do đó, hoàn tất quá trình xây dựng node.

Phân chia các dữ liệu theo cây quyết định đã xây dựng, ta được:

A group of rectangular colored rectangles

Description automatically generated

Bài 2: Dùng gini để vẽ cây quyết định với dataset cho trong file hw\_dataset\_2.csv, biến mục tiêu là OK?. Lưu ý Price là thuộc tính dạng liên tục.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of restaurant** | **Price** | **Neighborhood** | **Restriction** | **OK?** |
| Fast food | 5 | Oakland | Vegetarian | Not OK |
| Ethnic | 40 | Squirrel Hill | Gluten Free | Not OK |
| Casual dining | 35 | Squirrel Hill | None | Not OK |
| Casual dining | 80 | Shadyside | Vegetarian | Not OK |
| Casual dining | 11 | Oakland | Vegetarian | OK |
| Fast food | 31 | Squirrel Hill | None | OK |
| Ethnic | 10 | Squirrel Hill | None | OK |
| Casual dining | 22 | Shadyside | Gluten Free | Not OK |
| Fast food | 140 | Oakland | None | Not OK |
| Ethnic | 68 | Shadyside | Vegetarian | OK |
| Casual dining | 57 | Shadyside | Gluten Free | OK |

Cây quyết định được vẽ qua thư viện sklearn và visualize bằng thư viện sklearn.tree

A diagram of a number of equations

Description automatically generated with medium confidence