**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN BÁO CÁO THỰC HÀNH**

KHOA HỆ THỐNG THÔNG TIN Môn: Khai thác dữ liệu

**BÀI THỰC HÀNH SỐ 1**

Giảng viên hướng dẫn: ThS. Nguyễn Hồ Duy Trí

Sinh viên thực hiện: Phạm Thiện Bảo

MSSV: 20521107

Lớp: IS252.N22.HTCL

**Bài làm**

# I. Thực hiện lại phần hướng dẫn chung:

**1. Yêu cầu cơ bản:**

**Câu 1: Xác định loại của các thuộc tính trong bảng dữ liệu**

+ Thuộc tính Survival: là thuộc tính định tính thuộc loại nhị phân bất đối xứng (Asymmetric - binary).

+ Thuộc tính Pclass: là thuộc tính định tính thuộc loại xếp hạng (ordinal).

+ Thuộc tính Name: là thuộc tính định tính thuộc loại phân loại (nominal) .

+ Thuộc tính Sex: là thuộc tính định tính thuộc loại nhị phân đối xứng (Symmetric - binary).

+ Thuộc tính Age: là thuộc tính định lượng thuộc loại giá trị theo tỷ lệ (ratio-scaled).

+ Thuộc tính SibSp: là thuộc tính định lượng thuộc loại giá trị theo tỷ lệ (ratio-scaled).

+ Thuộc tính Parch: là thuộc tính định lượng thuộc loại giá trị theo tỷ lệ (ratio-scaled).

+ Thuộc tính Embarked: là thuộc tính định tính thuộc loại phân loại (nominal).

**Câu 2: Xử lý dữ liệu bị thiếu trong bảng dữ liệu**

Nhận thấy ở một số mẫu trong dữ liệu, tại thuộc tính Age có giá trị bị thiếu => Ta tiến hành điền vào phần khuyết bằng giá trị trung bình của các mẫu cùng thuộc một thuộc tính quyết định.

+ Xét trường hợp thuộc tính quyết định survival là 1, ta có:

Trung bình sống sót =

Trung bình chết =

Sau đó, ta lấy hai giá trị trung bình trên điền vào vị trí dữ liệu còn thiếu ở thuộc tính tuổi có survival là 1 và 0 tương ứng.

**Câu 3: Thực hiện khử nhiễu trên thuộc tính tuổi (Age) bằng kĩ thuật Binning và làm trơn (smoothing)**

+ Để binning (chia giỏ) ta thấy giá trị thấp nhất là 2 tuổi, lớn nhất là 58 tuổi nên ta chọn 0 và 60 là giá trị bắt đầu và kết thúc cho mỗi giỏ. Chọn độ rộng mỗi giỏ là 20 ta được các giỏ sau: (0, 20], (20, 40], (40, 60].

+ Để smoothing ta tính giá trị trung bình của các giá trị trong cùng một giỏ, sau đó thay thế giá trị tuổi bằng giá trị trung bình vừa tính theo mỗi giỏ.

Trung bình Age = (0, 20] =

Trung bình Age = (20, 40] =

Trung bình Age = (40, 60] =

**Câu 4: Thực hiện rời rạc hóa dữ liệu trên thuộc tính tuổi (Age) thay thế bởi:**

Ta thấy giá trị thấp nhất là 2 tuổi, lớn nhất là 58 tuổi nên rời rạc hóa như sau:

|  |  |
| --- | --- |
| Khoảng tuổi | Giá trị rời rạc |
| (0,12] | Children |
| (12,20] | Adolescence |
| (20,40] | Adult |
| (40, 58] | Senior |

**Câu 5: Biểu diễn khác của dạng dữ liệu categorical:**

Để đảm bảo công bằng cho các thuộc tính định tính, ta có thể biểu diễn thuộc tính này thành dạng One-hot Encoding. Ở dạng biểu diễn này, mỗi giá trị của thuộc tính được biểu diễn bằng một vector với một thành phần có giá trị là 1 và các thành phần còn lại có giá trị 0. Số lượng thành phần của vector chính là số loại của thuộc tính.

Ví dụ: Xét thuộc tính **“Cổng lên tàu”** (Embarked) trong bảng dữ liệu có 3 giá trị C, Q, S. Ta có thể biểu diễn One-hot như sau:

|  |  |
| --- | --- |
| **Embarked** | **One-hot encoding** |
| C | [1,0,0] |
| Q | [0,1,0] |
| S | [0,0,1] |

**Câu 6: Thực hiện chuẩn hóa dữ liệu trên bằng Min-max normalization**

Xét giá trị 7.25 của thuộc tính giá vé (Fare) áp dụng công thức tính thì ta được giá trị mới:

**2. Yêu cầu lập trình:**

Trình bày trong file *Lab1\_20521107.ipynb* đính kèm.

# II. Thực hiện khảo sát và tiền xử lý dữ liệu tập dữ liệu bank marketing:

Trình bày trong file *Lab1\_20521107.ipynb* đính kèm.