Đại học quốc gia TP HCM

Đại học Công nghệ thông tin

Khai thác dữ liệu

Thực hành Lab 1

GV: Nguyễn Hồ Duy Trí

Sinh viên: Nguyễn Hữu Thắng

MSSV: 19522208

Lớp: IS252.M22.1

Mục lục

[**I. BÀI 1 - HƯỚNG DẪN CHUNG: TITANIC** 2](#_Toc98435333)

[**1. Xác định loại của các thuộc tính trong bảng dữ liệu (nominal, categorical, binary...)** 2](#_Toc98435334)

[**2. Xử lý dữ liệu bị thiếu trong bảng dữ liệu** 2](#_Toc98435335)

[**3. Thực hiện khử nhiễu trên thuộc tính tuổi (Age) bằng kĩ thuật Binning và làm trơn (smoothing).** 4](#_Toc98435336)

[**4. Thực hiện rời rạc hóa dữ liệu trên thuộc tính tuổi (Age) thay thế bởi:** 5](#_Toc98435337)

[**5. Xét các thuộc tính dạng categorical, nhầm tránh biểu diễn sai giá trị thuộc tính khi sử dụng đối với một số thuật toán khác thác dữ liệu.** 7](#_Toc98435338)

[**6. Chuẩn hóa dữ liệu bằng Min – max normalization** 9](#_Toc98435339)

[**II. BÀI 2: BANK MARKETING** 11](#_Toc98435340)

[**1. Xác định loại của các thuộc tính trong bảng dữ liệu (nominal, categorical, binary...)** 11](#_Toc98435341)

[**2. Xử lý dữ liệu bị thiếu trong bảng dữ liệu** 11](#_Toc98435342)

[**3. Thực hiện khử nhiễu trên thuộc tính thời lượng tương tác cuối cùng bằng kỹ thuật binning và trơn** 13](#_Toc98435343)

[**4. Thực hiện rời rạc hóa dữ liệu thời lượng tương tác cuối cùng (Duration) thay thế bởi** 14](#_Toc98435344)

[**5. Xét các thuộc tính dạng categorical, nhầm tránh biểu diễn sai giá trị thuộc tính khi sử dụng đối với một số thuật toán khác thác dữ liệu.** 16](#_Toc98435345)

# **I. BÀI 1 - HƯỚNG DẪN CHUNG: TITANIC**

## **1. Xác định loại của các thuộc tính trong bảng dữ liệu (nominal, categorical, binary...)**

* **Survival**: Hành khách sống sót hay không (1: Yes/sống, 0: No/chết). Survival là thuộc tính quyết định (thuộc nhóm định tính – binary)
* **Pclass**: Loại vé (1: 1st, 2: 2nd, 3: 3rd) (thuộc nhóm định tính – ordinal)
* **Name**: Tên (thuộc nhóm định tính – Nominal)
* **Sex**: Giới tính (thuộc nhóm định tính – Binary)
* **Age**: Tuổi (thuộc nhóm định lượng – Interval-scaled)
* **SibSp**: Số vợ hoặc chồng/ anh chị em trên tàu (thuộc nhóm định lượng – Interval-scaled)
* **Fare**: Giá vé tàu (thuộc nhóm định lượng – Ratio-scaled)
* **Parch**: Số lượng cha mẹ/ con cái trên tàu (thuộc tính định lượng – Interval-scaled)
* **Embarked**: Cổng lên tàu (thuộc nhóm định tính – Ordinal)

## **2. Xử lý dữ liệu bị thiếu trong bảng dữ liệu**

* Dữ liệu thiếu trong cột Age
* Tính tuổi trung bình

*age\_mean = (22+38+26+35+35+54+2+27+14+4+58+20+39+14+55+2+31)/17*

*= 28 (tuổi)*

* Điền giá trị trung bình vào các ô còn thiếu ở hàng thứ 6, hàng 18 và hàng 20 của cột Age.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Survival | Pclass | Name | Sex | Age | SibSp | Parch | Fare | Embarked |
| 0 | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | 7.25 | S |
| 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 38 | 1 | 0 | 71.2833 | C |
| 1 | 3 | Heikkinen, Miss. Laina | female | 26 | 0 | 0 | 7.925 | S |
| 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35 | 1 | 0 | 53.1 | S |
| 0 | 3 | Allen, Mr. William Henry | male | 35 | 0 | 0 | 8.05 | S |
| 0 | 3 | Moran, Mr. James | male | **28** | 0 | 0 | 8.4583 | Q |
| 0 | 1 | McCarthy, Mr. Timothy J | male | 54 | 0 | 0 | 51.8625 | S |
| 0 | 3 | Palsson, Master. Gosta Leonard | male | 2 | 3 | 1 | 21.075 | S |
| 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 27 | 0 | 2 | 11.1333 | S |
| 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 14 | 1 | 0 | 30.0708 | C |
| 1 | 3 | Sandstrom, Miss. Marguerite Rut | female | 4 | 1 | 1 | 16.7 | S |
| 1 | 1 | Bonnell, Miss. Elizabeth | female | 58 | 0 | 0 | 26.55 | S |
| 0 | 3 | Saundercock, Mr. William Henry | male | 20 | 0 | 0 | 8.05 | S |
| 0 | 3 | Andersson, Mr. Anders Johan | male | 39 | 1 | 5 | 31.275 | S |
| 0 | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 14 | 0 | 0 | 7.8542 | S |
| 1 | 2 | Hewlett, Mrs. (Mary D Kingcome) | female | 55 | 0 | 0 | 16 | S |
| 0 | 3 | Rice, Master. Eugene | male | 2 | 4 | 1 | 29.125 | Q |
| 1 | 2 | Williams, Mr. Charles Eugene | male | **28** | 0 | 0 | 13 | S |
| 0 | 3 | Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) | female | 31 | 1 | 0 | 18 | S |
| 1 | 3 | Masselmani, Mrs. Fatima | female | **28** | 0 | 0 | 7.225 | C |

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

* Để binning thì 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 cho mỗi giỏ là 20, ta được 3 giỏ sau: (0, 20], (20, 40], (40, 60]
* Đem dữ liệu tuổi thêm vào các giỏ
* Giỏ 1: 2, 2, 4, 14, 20
* Giỏ 2: 22, 26, 27, 28, 28, 28, 31, 35, 35, 38, 39
* Giỏ 3: 54, 55, 58
* Làm mịn bằng biên giỏ
* Giỏ 1: 2, 2, 2, 20, 20
* Giỏ 2: 22, 22, 22, 22, 22, 22, 39, 39, 39, 39, 39
* Giỏ 3: 54, 54, 58
* Thay thế giá trị tuổi bằng giá trị vừa tính theo mỗi giỏ

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Survival | Pclass | Name | Sex | Age | SibSp | Parch | | Fare | Embarked |
| 0 | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | 7.25 | | S |
| 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 39 | 1 | 0 | 71.2833 | | C |
| 1 | 3 | Heikkinen, Miss. Laina | female | 22 | 0 | 0 | 7.925 | | S |
| 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 39 | 1 | 0 | 53.1 | | S |
| 0 | 3 | Allen, Mr. William Henry | male | 39 | 0 | 0 | 8.05 | | S |
| 0 | 3 | Moran, Mr. James | male | 22 | 0 | 0 | 8.4583 | | Q |
| 0 | 1 | McCarthy, Mr. Timothy J | male | 54 | 0 | 0 | 51.8625 | | S |
| 0 | 3 | Palsson, Master. Gosta Leonard | male | 2 | 3 | 1 | 21.075 | | S |
| 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 22 | 0 | 2 | 11.1333 | | S |
| 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 20 | 1 | 0 | 30.0708 | | C |
| 1 | 3 | Sandstrom, Miss. Marguerite Rut | female | 2 | 1 | 1 | 16.7 | | S |
| 1 | 1 | Bonnell, Miss. Elizabeth | female | 58 | 0 | 0 | 26.55 | | S |
| 0 | 3 | Saundercock, Mr. William Henry | male | 20 | 0 | 0 | 8.05 | | S |
| 0 | 3 | Andersson, Mr. Anders Johan | male | 39 | 1 | 5 | 31.275 | | S |
| 0 | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 20 | 0 | 0 | 7.8542 | | S |
| 1 | 2 | Hewlett, Mrs. (Mary D Kingcome) | female | 54 | 0 | 0 | 16 | | S |
| 0 | 3 | Rice, Master. Eugene | male | 2 | 4 | 1 | 29.125 | | Q |
| 1 | 2 | Williams, Mr. Charles Eugene | male | 22 | 0 | 0 | 13 | | S |
| 0 | 3 | Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) | female | 39 | 1 | 0 | 18 | | S |
| 1 | 3 | Masselmani, Mrs. Fatima | female | 22 | 0 | 0 | 7.225 | | C |

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

* Khoảng giá trị (10 – 20, 0 – 10…)
* Bằng các nhãn khái niệm (youth, senior, adult…)
* Senior: > 55
* Adult: [36, 55]
* Youth: < 18
* Young Adult: [18, 35]

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Survival | Pclass | Name | Sex | Age New | Age Group | SibSp | Parch | Fare | Embarked |
| 0 | 3 | Braund, Mr. Owen Harris | male | 20-29 | Young Adult | 1 | 0 | 7.25 | S |
| 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 30-39 | Adult | 1 | 0 | 71.2833 | C |
| 1 | 3 | Heikkinen, Miss. Laina | female | 20-29 | Young Adult | 0 | 0 | 7.925 | S |
| 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 30-39 | Adult | 1 | 0 | 53.1 | S |
| 0 | 3 | Allen, Mr. William Henry | male | 30-39 | Adult | 0 | 0 | 8.05 | S |
| 0 | 3 | Moran, Mr. James | male | 20-29 | Young Adult | 0 | 0 | 8.4583 | Q |
| 0 | 1 | McCarthy, Mr. Timothy J | male | 50-59 | Adult | 0 | 0 | 51.8625 | S |
| 0 | 3 | Palsson, Master. Gosta Leonard | male | 0-9 | Youth | 3 | 1 | 21.075 | S |
| 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 20-29 | Young Adult | 0 | 2 | 11.1333 | S |
| 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 20-29 | Young Adult | 1 | 0 | 30.0708 | C |
| 1 | 3 | Sandstrom, Miss. Marguerite Rut | female | 0-9 | Youth | 1 | 1 | 16.7 | S |
| 1 | 1 | Bonnell, Miss. Elizabeth | female | 50-59 | Senior | 0 | 0 | 26.55 | S |
| 0 | 3 | Saundercock, Mr. William Henry | male | 20-29 | Young Adult | 0 | 0 | 8.05 | S |
| 0 | 3 | Andersson, Mr. Anders Johan | male | 30-39 | Adult | 1 | 5 | 31.275 | S |
| 0 | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 20-29 | Young Adult | 0 | 0 | 7.8542 | S |
| 1 | 2 | Hewlett, Mrs. (Mary D Kingcome) | female | 50-59 | Adult | 0 | 0 | 16 | S |
| 0 | 3 | Rice, Master. Eugene | male | 0-9 | Youth | 4 | 1 | 29.125 | Q |
| 1 | 2 | Williams, Mr. Charles Eugene | male | 20-29 | Young Adult | 0 | 0 | 13 | S |
| 0 | 3 | Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) | female | 30-39 | Adult | 1 | 0 | 18 | S |
| 1 | 3 | Masselmani, Mrs. Fatima | female | 20-29 | Young Adult | 0 | 0 | 7.225 | C |

## **5. Xét các thuộc tính dạng categorical, nhầm tránh biểu diễn sai giá trị thuộc tính khi sử dụng đối với một số thuật toán khác thác dữ liệu.**

* Để đảm bảo công bằng cho các thuộc tính, ta biểu diễn thuộc tính này thành dạng One-hot Encoding. 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.
* Ta biểu diễn như sau: S -> 1,0,0; C ->0,1,0; Q -> 0,0,1.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Survival | Pclass | Name | Sex | Age New | Age Group | SibSp | Parch | Fare | Embarked |
| 0 | 3 | Braund, Mr. Owen Harris | male | 20-29 | Young Adult | 1 | 0 | 7.25 | 1,0,0 |
| 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 30-39 | Adult | 1 | 0 | 71.2833 | 0,1,0 |
| 1 | 3 | Heikkinen, Miss. Laina | female | 20-29 | Young Adult | 0 | 0 | 7.925 | 1,0,0 |
| 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 30-39 | Adult | 1 | 0 | 53.1 | 1,0,0 |
| 0 | 3 | Allen, Mr. William Henry | male | 30-39 | Adult | 0 | 0 | 8.05 | 1,0,0 |
| 0 | 3 | Moran, Mr. James | male | 20-29 | Young Adult | 0 | 0 | 8.4583 | 0,0,1 |
| 0 | 1 | McCarthy, Mr. Timothy J | male | 50-59 | Adult | 0 | 0 | 51.8625 | 1,0,0 |
| 0 | 3 | Palsson, Master. Gosta Leonard | male | 0-9 | Youth | 3 | 1 | 21.075 | 1,0,0 |
| 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 20-29 | Young Adult | 0 | 2 | 11.1333 | 1,0,0 |
| 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 20-29 | Young Adult | 1 | 0 | 30.0708 | 0,1,0 |
| 1 | 3 | Sandstrom, Miss. Marguerite Rut | female | 0-9 | Youth | 1 | 1 | 16.7 | 1,0,0 |
| 1 | 1 | Bonnell, Miss. Elizabeth | female | 50-59 | Senior | 0 | 0 | 26.55 | 1,0,0 |
| 0 | 3 | Saundercock, Mr. William Henry | male | 20-29 | Young Adult | 0 | 0 | 8.05 | 1,0,0 |
| 0 | 3 | Andersson, Mr. Anders Johan | male | 30-39 | Adult | 1 | 5 | 31.275 | 1,0,0 |
| 0 | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 20-29 | Young Adult | 0 | 0 | 7.8542 | 1,0,0 |
| 1 | 2 | Hewlett, Mrs. (Mary D Kingcome) | female | 50-59 | Adult | 0 | 0 | 16 | 1,0,0 |
| 0 | 3 | Rice, Master. Eugene | male | 0-9 | Youth | 4 | 1 | 29.125 | 0,0,1 |
| 1 | 2 | Williams, Mr. Charles Eugene | male | 20-29 | Young Adult | 0 | 0 | 13 | 1,0,0 |
| 0 | 3 | Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) | female | 30-39 | Adult | 1 | 0 | 18 | 1,0,0 |
| 1 | 3 | Masselmani, Mrs. Fatima | female | 20-29 | Young Adult | 0 | 0 | 7.225 | 0,1,0 |

## **6. Chuẩn hóa dữ liệu bằng Min – max normalization**

* Ta chuyển dữ liệu về khoảng giá trị thuộc vùng 0 – 1 bằng công thức sau:

new\_value =

Trong đó:

* New\_value = giá trị sau khi chuyển đổi
* Value = giá trị đang cần chuyển đổi
* Min = giá trị thấp nhất của thuộc tính
* Max = giá trị lớn nhất của thuộc tính

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Survival | Pclass | Name | Sex | Age New | Age Group | SibSp | Parch | Fare | Embarked |
| 0 | 3 | Braund, Mr. Owen Harris | male | 20-29 | Young Adult | 1 | 0 | 0.000390269 | 1,0,0 |
| 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 30-39 | Adult | 1 | 0 | 1 | 0,1,0 |
| 1 | 3 | Heikkinen, Miss. Laina | female | 20-29 | Young Adult | 0 | 0 | 0.010927546 | 1,0,0 |
| 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 30-39 | Adult | 1 | 0 | 0.716144512 | 1,0,0 |
| 0 | 3 | Allen, Mr. William Henry | male | 30-39 | Adult | 0 | 0 | 0.012878893 | 1,0,0 |
| 0 | 3 | Moran, Mr. James | male | 20-29 | Young Adult | 0 | 0 | 0.019252774 | 0,0,1 |
| 0 | 1 | McCarthy, Mr. Timothy J | male | 50-59 | Adult | 0 | 0 | 0.696826172 | 1,0,0 |
| 0 | 3 | Palsson, Master. Gosta Leonard | male | 0-9 | Youth | 3 | 1 | 0.216209297 | 1,0,0 |
| 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 20-29 | Young Adult | 0 | 2 | 0.06101161 | 1,0,0 |
| 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 20-29 | Young Adult | 1 | 0 | 0.356640748 | 0,1,0 |
| 1 | 3 | Sandstrom, Miss. Marguerite Rut | female | 0-9 | Youth | 1 | 1 | 0.147912136 | 1,0,0 |
| 1 | 1 | Bonnell, Miss. Elizabeth | female | 50-59 | Senior | 0 | 0 | 0.301678315 | 1,0,0 |
| 0 | 3 | Saundercock, Mr. William Henry | male | 20-29 | Young Adult | 0 | 0 | 0.012878893 | 1,0,0 |
| 0 | 3 | Andersson, Mr. Anders Johan | male | 30-39 | Adult | 1 | 5 | 0.375439248 | 1,0,0 |
| 0 | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 20-29 | Young Adult | 0 | 0 | 0.009822302 | 1,0,0 |
| 1 | 2 | Hewlett, Mrs. (Mary D Kingcome) | female | 50-59 | Adult | 0 | 0 | 0.136984591 | 1,0,0 |
| 0 | 3 | Rice, Master. Eugene | male | 0-9 | Youth | 4 | 1 | 0.341876072 | 0,0,1 |
| 1 | 2 | Williams, Mr. Charles Eugene | male | 20-29 | Young Adult | 0 | 0 | 0.090152252 | 1,0,0 |
| 0 | 3 | Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) | female | 30-39 | Adult | 1 | 0 | 0.16820615 | 1,0,0 |
| 1 | 3 | Masselmani, Mrs. Fatima | female | 20-29 | Young Adult | 0 | 0 | 0 | 0,1,0 |

# **II. BÀI 2: BANK MARKETING**

## **1. Xác định loại của các thuộc tính trong bảng dữ liệu (nominal, categorical, binary...)**

* Job: Nghề nghiệp (thuộc nhóm định tính – Nominal)
* Marial: Tình trạng hôn nhân (thuộc nhóm định tính - Nominal)
* Education: trình độ học vấn (thuộc nhóm định tính – Ordinal)
* Default: Đã có tính dụng mặc định hay chưa (thuộc nhóm định tính – Binary-Asymmetric)
* Duration: Thời lượng lần tương tác cuối cùng, tính trên đơn vị giây (thuộc nhóm định lượng – Ratio-scaled)
* Cons.price.idx: Chỉ số tiêu dùng (thuộc nhóm định lượng – Interval-scaled)
* Subcribe: Theo dõi (Yes/ No thuộc nhóm định tính – Binary-Asymmetric)

## **2. Xử lý dữ liệu bị thiếu trong bảng dữ liệu**

Dữ liệu bị thiếu: Marial và Cons.price.idx

* Điền mode vào cột dữ liệu Marial
* Tính tần số xuất hiện: Married 10 lần, Divorced 2 lần, Single 3 lần => mode là Married
* Điền Married vào chỗ còn thiếu của cột Marial
* Điền giá trị trung bình vào cột dữ liệu Cons.price.idx
* Trungbinh = 98.83
* Điền 98.83 vào những ô còn thiếu của cột Cons.price.idx

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Job** | | | **Marial** | **Education** | **Default** | **Loan** | **Duration** | **Cons.price.idx** | | **Subscribe** | |
| housemaid | Married | | basic.4y | no | no | 261 | 93.994 | no | |
| services | Married | | high.school | unknown | no | 149 | 93.994 | no | |
| blue-collar | Divorced | | basic.4y | unknown | no | 1575 | 93.994 | yes | |
| admin. | Single | | high.school | no | no | 338 | 93.994 | no | |
| blue-collar | Single | | basic.9y | no | no | 179 | 93.994 | no | |
| services | Married | | high.school | no | no | 1030 | 93.994 | yes | |
| management | Married | | high.school | unknown | no | 149 | **93.83** | no | |
| unemployed | **Married** | | university.degree | unknown | no | 424 | 93.994 | no | |
| technician | Married | | professional.course | no | no | 1623 | 93.994 | yes | |
| services | Divorced | | high.school | unknown | no | 568 | 93.994 | no | |
| blue-collar | Married | | high.school | no | no | 1297 | 93.994 | yes | |
| self-employed | Married | | basic.9y | no | no | 376 | 93.918 | no | |
| entrepreneur | Married | | professional.course | no | yes | 576 | 93.994 | no | |
| services | Single | | high.school | no | no | 1059 | 93.2 | yes | |
| technician | Married | | basic.9y | no | no | 705 | 93.2 | yes | |
| Retired | Married | | high.school | no | no | 532 | 93.2 | no | |

## **3. Thực hiện khử nhiễu trên thuộc tính thời lượng tương tác cuối cùng bằng kỹ thuật binning và trơn**

* Sắp xếp các Duration theo thứ tự tăng dần: 149, 149, 179, 261, 338, 376, 424, 532, 568, 576, 705, 1030, 1059, 1297, 1575, 1623.
* Chia thành 4 giỏ

Giỏ 1: 149, 149, 179, 261

Giỏ 2: 338, 376, 424, 532

Giỏ 3: 568, 576, 705, 1030

Giỏ 4: 1059, 1297, 1575, 1623

* Làm mịn bằng biên giỏ

Giỏ 1: 149, 149, 149, 261

Giỏ 2: 338, 338, 338, 532

Giỏ 3: 568, 568, 568, 1030

Giỏ 4: 1059, 1059, 1623, 1623

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Job** | | **Marial** | **Education** | **Default** | **Loan** | **Duration** | **Cons.price.idx** | **Subscribe** |
| housemaid | Married | basic.4y | no | no | 261 | 93.994 | no |
| services | Married | high.school | unknown | no | 149 | 93.994 | no |
| blue-collar | Divorced | basic.4y | unknown | no | 1623 | 93.994 | yes |
| admin. | Single | high.school | no | no | 338 | 93.994 | no |
| blue-collar | Single | basic.9y | no | no | 149 | 93.994 | no |
| services | Married | high.school | no | no | 1030 | 93.994 | yes |
| management | Married | high.school | unknown | no | 149 | 93.83 | no |
| unemployed | Married | university.degree | unknown | no | 338 | 93.994 | no |
| technician | Married | professional.course | no | no | 1623 | 93.994 | yes |
| services | Divorced | high.school | unknown | no | 568 | 93.994 | no |
| blue-collar | Married | high.school | no | no | 1059 | 93.994 | yes |
| self-employed | Married | basic.9y | no | no | 338 | 93.918 | no |
| entrepreneur | Married | professional.course | no | yes | 568 | 93.994 | no |
| services | Single | high.school | no | no | 1059 | 93.2 | yes |
| technician | Married | basic.9y | no | no | 568 | 93.2 | yes |
| Retired | Married | high.school | no | no | 532 | 93.2 | no |

## **4. Thực hiện rời rạc hóa dữ liệu thời lượng tương tác cuối cùng (Duration) thay thế bởi**

* Khoảng giá trị (0-500, 501-1000, 1001-1500, 1501-2000)
* Bằng các nhãn khái niệm (Many, Few, …)

Few < 1000

Many >= 1000

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Job** | | **Marial** | **Education** | **Default** | **Loan** | **Duration new** | **Duration group** | **Cons.price.idx** | **Subscribe** |
| housemaid | Married | basic.4y | no | no | 0-500 | Few | 93.994 | no |
| services | Married | high.school | unknown | no | 0-500 | Few | 93.994 | no |
| blue-collar | Divorced | basic.4y | unknown | no | 1501-2000 | Many | 93.994 | yes |
| admin. | Single | high.school | no | no | 0-500 | Few | 93.994 | no |
| blue-collar | Single | basic.9y | no | no | 0-500 | Few | 93.994 | no |
| services | Married | high.school | no | no | 1001-1500 | Many | 93.994 | yes |
| management | Married | high.school | unknown | no | 0-500 | Few | 93.83 | no |
| unemployed | Married | university.degree | unknown | no | 0-500 | Few | 93.994 | no |
| technician | Married | professional.course | no | no | 1501-2000 | Many | 93.994 | yes |
| services | Divorced | high.school | unknown | no | 501-1000 | Few | 93.994 | no |
| blue-collar | Married | high.school | no | no | 1001-1500 | Many | 93.994 | yes |
| self-employed | Married | basic.9y | no | no | 0-500 | Few | 93.918 | no |
| entrepreneur | Married | professional.course | no | yes | 501-1000 | Few | 93.994 | no |
| services | Single | high.school | no | no | 1001-1500 | Many | 93.2 | yes |
| technician | Married | basic.9y | no | no | 501-1000 | Few | 93.2 | yes |
| Retired | Married | high.school | no | no | 501-1000 | Few | 93.2 | no |

## **5. Xét các thuộc tính dạng categorical, nhầm tránh biểu diễn sai giá trị thuộc tính khi sử dụng đối với một số thuật toán khác thác dữ liệu.**

Dùng định dạng One-hot Encoding để biểu diễn

* Ta biểu diễn như sau: Married -> 1,0,0; Divorced -> 0,1,0; Single -> 0,0,1

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Job** | | **Marial** | **Education** | **Default** | **Loan** | **Duration new** | **Duration group** | **Cons.price.idx** | | **Subscribe** |
| housemaid | 1,0,0 | basic.4y | no | no | 0-500 | Few | | 93.994 | no |
| services | 1,0,0 | high.school | unknown | no | 0-500 | Few | | 93.994 | no |
| blue-collar | 0,1,0 | basic.4y | unknown | no | 1501-2000 | Many | | 93.994 | yes |
| admin. | 0,0,1 | high.school | no | no | 0-500 | Few | | 93.994 | no |
| blue-collar | 0,0,1 | basic.9y | no | no | 0-500 | Few | | 93.994 | no |
| services | 1,0,0 | high.school | no | no | 1001-1500 | Many | | 93.994 | yes |
| management | 1,0,0 | high.school | unknown | no | 0-500 | Few | | 93.83 | no |
| unemployed | 1,0,0 | university.degree | unknown | no | 0-500 | Few | | 93.994 | no |
| technician | 1,0,0 | professional.course | no | no | 1501-2000 | Many | | 93.994 | yes |
| services | 0,1,0 | high.school | unknown | no | 501-1000 | Few | | 93.994 | no |
| blue-collar | 1,0,0 | high.school | no | no | 1001-1500 | Many | | 93.994 | yes |
| self-employed | 1,0,0 | basic.9y | no | no | 0-500 | Few | | 93.918 | no |
| entrepreneur | 1,0,0 | professional.course | no | yes | 501-1000 | Few | | 93.994 | no |
| services | 0,0,1 | high.school | no | no | 1001-1500 | Many | | 93.2 | yes |
| technician | 1,0,0 | basic.9y | no | no | 501-1000 | Few | | 93.2 | yes |
| Retired | 1,0,0 | high.school | no | no | 501-1000 | Few | | 93.2 | no |