# 3. Trích xuất đặc trưng

## 3.1. Xử lí dữ liệu ban đầu

Với dữ liệu đầu vào gồm 64461 hàng × 61 cột, nhóm tiến hành giảm chiều dữ liệu bằng cách loại bỏ những hàng có quốc gia không nằm ở châu Á. Đồng thời, dataset chỉ giữ lại những trường (field) liên quan, đó là 5 trường: “ConvertedComp”, “Country”, “DevType”, “YearsCode”, “EdLevel”. Kết quả thu được là bộ dữ liệu con gồm 16278 hàng × 5 cột:

Graphical user interface, text, application

Description automatically generated

Hình . Kết quả xử lý dữ liệu ban đầu.

## 3.2. Làm sạch dữ liệu

Nhóm làm sạch dữ liệu bằng cách xoá các quan sát có bất kì phần tử dữ liệu nào bị bỏ trống (xoá các phần tử NaN). Kết quả thu được bộ dữ liệu bị giảm số hàng còn 5645 hàng:

Graphical user interface, text, application, email

Description automatically generated

Hình 2. Kết quả xử lý dữ liệu đã được làm sạch.

## 3.3. Xử lí trường “DevType”

Trong bộ dữ liệu, trường “DevType” có kiểu dữ liệu là xâu (string) tuy nhiên giá trị của nó lại có dạng danh sách các xâu cách nhau bởi dấu chấm phẩy (‘;’). Ví dụ:

Graphical user interface, text, application

Description automatically generated

Hình 3. Trường “DevType” trước khi được xử lí.

Nhằm mã hóa trường “DevType”, Nhóm sử dụng mô hình Bag of Words Model. Ý tưởng của thuật toán là biểu diễn cho mỗi mẫu dữ liệu văn bản dưới dạng một vector số, trong đó mỗi chiều là một từ cụ thể trong kho dữ liệu và giá trị có thể là tần số của nó xuất hiện trong đoạn văn bản (giá trị có thể là 0 hoặc 1).

Table

Description automatically generated

Hình 4. Kết quả mã hóa trường “DevType”.

Sau đó, chuyển kết quả thu được thành cấu trúc dữ liệu dataframe, đồng thời giảm chiều (số hàng) bằng cách chỉ lấy những trường bắt đầu bằng “Developer”. Kết quả: Table

Description automatically generated

Hình 5. Kết quả bước xử lý trường “DevType”.

Cuối cùng, lọc dataset ban đầu dựa vào chỉ số hàng (index) của dataframe đã được xử lí. Mục đích của việc này là để index của các dataframe của các trường khác sẽ trùng khớp với nhau. Kết quả thu được:

Graphical user interface, text, application

Description automatically generated

Hình 6. Kết quả lọc dataset ban đầu dựa vào index của dataframe vừa xử lý.

## 3.4. Xử lí trường “Country”

Nhóm mã hóa trường “Country” có kiểu dữ liệu xâu (string) trong bộ dữ liệu bằng cách sử dụng kĩ thuật mã hoá One-hot encoding. Cụ thể, One-hot encoding là quá trình biến đổi từng giá trị thành các đặc trưng nhị phân chỉ chứa giá trị 1 hoặc 0. Mỗi mẫu trong đặc trưng phân loại sẽ được biến đổi thành một vecto có kích thước m chỉ với một trong các giá trị là 1 (biểu thị nó là active) . Dùng hàm fit() và transform() của lớp OneHotEncoder của thư viện sklearn.preprocessing. Sau đó, chuyển kết quả thu được thành cấu trúc dữ liệu dataframe:

Table

Description automatically generated

Hình 7. Trường “Country” trước khi được xử lí.

Table

Description automatically generated

Hình 8. Kết quả bước xử lý trường “Country”.

## 3.5. Xử lí trường “EdLevel”

Theo quan sát, nhóm thấy đây là dữ liệu phân loại theo thứ tự. Nhóm dùng map() của pandas để ánh xạ các đặc trưng này thành dạng số theo trình độ học vấn của người được phỏng vấn. Cách thức ánh xạ cụ thể được cho theo bảng bên dưới. Kết quả thu được được chuyển thành cấu trúc dữ liệu dataframe. Kết quả thu được:

Text, table

Description automatically generated

Hình 9. Trường “EdLevel” trước khi được xử lí.

Graphical user interface, application

Description automatically generated

Hình 10. Kết quả bước xử lý trường “EdLevel”.

|  |  |
| --- | --- |
| Giá trị ban đầu | Giá trị mã hóa |
| I never completed any formal education | 0 |
| Primary/elementary school | 1 |
| Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.) | 2 |
| Some college/university study without earning a degree | 3 |
| Associate degree (A.A., A.S., etc.) | 4 |
| Bachelor’s degree (B.A., B.S., B.Eng., etc.) | 5 |
| Master’s degree (M.A., M.S., M.Eng., MBA, etc.) | 6 |
| Other doctoral degree (Ph.D., Ed.D., etc.) | 7 |
| Professional degree (JD, MD, etc.) | 8 |

Bảng . Cách thức ánh xạ nhằm mã hóa trường “EdLevel”.

## 3.6. Xử lí trường “YearsCode”

Các phần tử dữ liệu có giá trị là “More than 50 years” và “Less than 1 year” sẽ lần lượt được chuyển thành 60 và 0.5. Chuyển kết quả thu được thành cấu trúc dữ liệu dataframe:

Table

Description automatically generated

Hình 11. Trường “YearsCode” trước khi được xử lí.

Table

Description automatically generated

Hình 12. Kết quả bước xử lý trường “YearsCode”.

## 3.7. Tạo bộ dữ liệu đã được xử lí

Nối các trường đã được xử lí ở trên với trường “ConvertedComp” ban đầu để tạo thành bộ dữ liệu hoàn chỉnh để đùng cho quá trình train. Kết quả dataset cuối cùng thu được:

**Table

Description automatically generated**

Hình 13. Bộ dữ liệu hoàn chỉnh đã xử lý.