CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

1. Môi trường thực nghiệm

* Google Colab là một công cụ mạnh mẽ để phát triển, thử nghiệm, và triển khai các mô hình học máy, bao gồm cả việc huấn luyện và dự đoán với mô hình hồi quy tuyến tính.

**A logo with yellow circles

Description automatically generated**

Hình 3.1. Google Colab

* Ngôn ngữ lập trình: Python 3.7 và các thư viện hỗ trợ khác như Sklearn, Numpy, Pandas, Matplotlib và Seaborn.

A logo of a software company

Description automatically generated

Hình 3.2. Ngôn ngữ lập trình Python

* Lưu trữ: Google Drive, dịch vụ lưu trữ đám mây, cho phép người dùng lưu trữ và truy cập các tệp tin trực tuyến.



Hình 3.3. Google Drive

1. Dữ liệu thực nghiệm

Bộ dữ liệu sử dụng trong nghiên cứu này được thu thập những người mà chính quyền tuyển dụng và cách trả lương cho nhân viên. Dữ liệu bao gồm tên, chức danh công việc và mức lương cho nhân viên thành phố San Francisco hầng năm từ năm 2011 đến năm 2014.

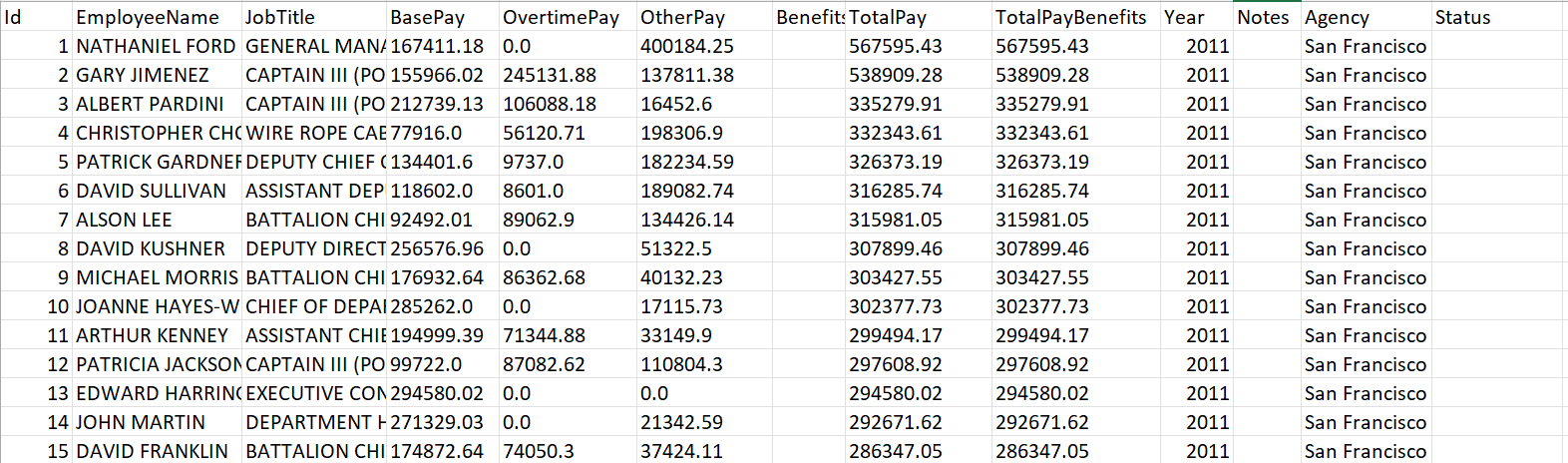
Bộ dữ liệu được phân tích ở đây là file dataset (.csv) có tên cụ thể là Salaries.csv, chứa 148654 bản ghi thông tin về thu thập của nhân viên.

Thông tin cụ thể như sau:

- Tên bộ dữ liệu: SF Salaries

- Nguồn:<https://www.kaggle.com/datasets/kaggle/sf-salaries>

- Dữ liệu 15 dòng đầu của dataset:

****

Hình 3.4. 15 dòng đầu của bộ dữ liệu gốc

**-** Thông tin cụ thể các cột của dataset như sau:

**Id**: Định danh gán cho nhân viên.

**EmployeeName**:Tên của nhân viên.

**JobTitle**: Chức danh của nhân viên.

**BasePay**: Lương cơ bản của nhân viên.

**OvertimePay**: tiền lương làm thêm giờ.

**OtherPay:** Các khoản thu thập khác ngoài lương cơ bản và làm thêm giờ.

**Benefits**: Các khoản phúc lợi do tổ chức cung cấp.

**TotalPay**: Tổng thu nhập trước khi tính thêm phúc lợi.

**TotalPayBenefits**: Tổng thu nhập sau khi thêm phúc lợi.

**Year**: Năm dữ liệu được ghi nhận.

**Notes:** Ghi chú liên quan đến nhân viên.

**Agency:** Tên tổ chức mà nhân viên làm việc.

**Status:** Tình trạng công việc hoặc trạng thái của nhân viên.

1. Quy Trình thực nghiệm

**A diagram of a company

Description automatically generated with medium confidence**

Hình 3.5. Quy trình thực nghiệm đề tài phân tích dữ liệu

1. Đặt mục tiêu

* Phân tích mô tả để thể hiện mối quan hệ giữa các giá trị của dữ liệu, từ đó đánh giá được tương quan của tổng thu nhập của nhân viên.
* Phân tích hồi quy để dự báo tổng thu nhập dựa theo mô hình hồi quy tuyến tính.

1. Tiền xử lý dữ liệu

* Làm sạch dữ liệu:

Làm sạch dữ liệu là quá trình loại bỏ các sai sót, lỗi, nhiễu và thông tin không chính xác hoặc không cần thiết khỏi tập dữ liệu ban đầu để đảm bảo dữ liệu đáng tin cậy và phù hợp cho việc phân tích và xử lý tiếp theo. Quá trình làm sạch dữ liệu thường là một phần quan trọng trong tiền xử lý dữ liệu trước khi bắt đầu phân tích mô tả và cả phân tích hồi quy.

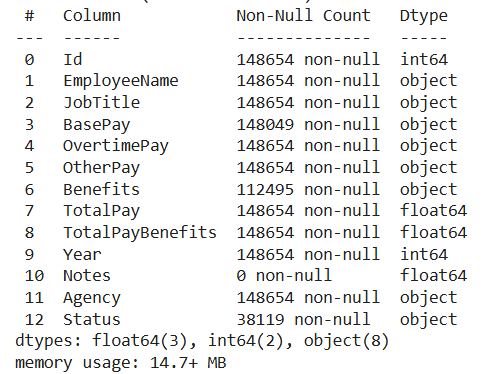
Một số tác vụ chính trong quá trình làm sạch dữ liệu bao gồm:

* Loại bỏ dữ liệu trùng lặp: Loại bỏ các bản ghi bị trùng lặp trong tập dữ liệu để tránh ảnh hưởng đến kết quả phân tích.
* Xử lý dữ liệu thiếu: Điền vào các giá trị thiếu hoặc quyết định loại bỏ chúng dựa trên ngữ cảnh và mục tiêu của phân tích.
* Xử lý giá trị ngoại lai : Xử lý giá trị ngoại lai bao gồm việc phát hiện và xử lý các điểm dữ liệu bất thường, nằm xa phân phối dữ liệu chính, bằng các phương pháp như loại bỏ, thay thế hoặc biến đổi
* Sửa lỗi và sai sót: Điều tra và sửa các lỗi cú pháp, sai sót chính tả hoặc sai sót logic trong dữ liệu.
* Chọn lọc đặc trưng: Xác định và lựa chọn các đặc trưng quan trọng nhất để sử dụng trong phân tích hoặc mô hình hóa.
* Xử lý cột số chứa giá trị phi số: chuyển đổi các cột có giá trị phi số sang giá trị số để tăng giá trị cho mô hình.

A computer code with text

Description automatically generated with medium confidence

Hình 3.6. Xử lý cột số chứa giá trị phi số



Hình 3.7. Trước khi xử lý cột số chứa giá trị phi số

Thông tin dữ liệu sau khi xử lý:

A screenshot of a computer

Description automatically generated

Hình 3.8. Sau khi xử lý cột số chứa giá trị phi số

* Kiểm tra dữ liệu khuyết :

A screenshot of a computer

Description automatically generated

Hình 3.9. Kiểm tra dữ liệu bị khuyết

* Ta sẽ điền khuyết bằng median :

A computer screen shot of a code

Description automatically generated

Hình 3.10. Xử lý dữ liệu bị khuyết

A screenshot of a phone

Description automatically generated

Hình 3.11. Sau khi xử lý dữ liệu bị khuyết

* Xử lý giá trị trùng lặp :



Hình 3.12. Xử lý dữ liệu bị trùng

A black and white text

Description automatically generated

Hình 3.13. Kết quả xử lý dữ liệu bị trùng

* Xử lý giá trị ngoại lai bằng phương pháp IQR

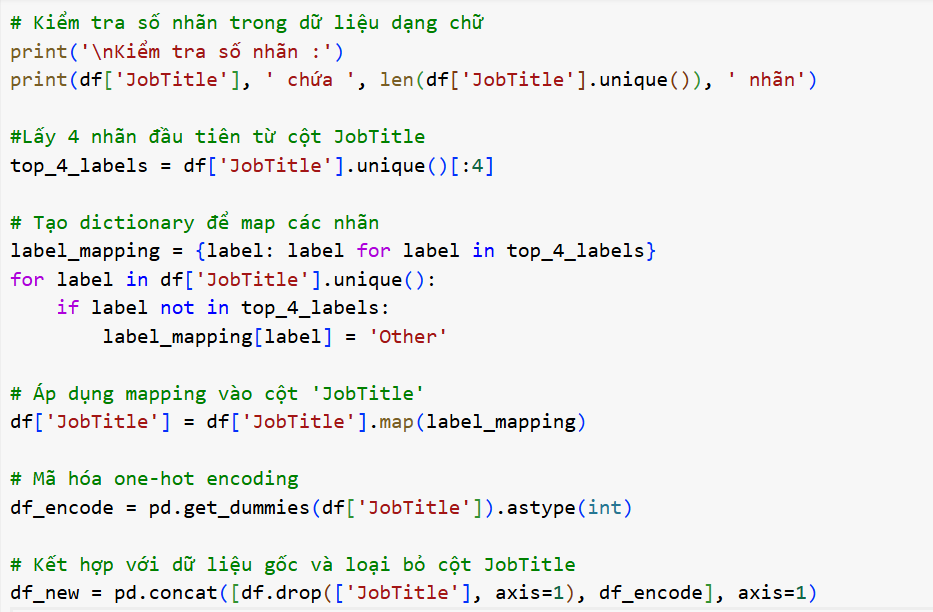
A screenshot of a computer program

Description automatically generated

Hình 3.14. Xử lý giá trị ngoại lai

Chuyển đổi dữ liệu: Chuyển đổi dữ liệu trong phân tích dữ liệu là quá trình thay đổi cách thức biểu diễn, xử lý hoặc áp dụng các phép toán trên dữ liệu ban đầu để tạo ra dữ liệu mới có ý nghĩa hoặc thuận tiện hơn cho mục đích phân tích. Nó có vai trò quan trọng trong việc biểu diễn trực quan hơn dataset, thuận tiện hơn trong việc phân tích dữ liệu.

Trong project này, ta thấy có cột duy nhất đang ở dạng Object, tức phi số, vì vậy, ta sẽ chuyển hóa cột dữ liệu này.



Hình 3.15. chuyển hóa cột dữ liệu