TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

KHOA CÔNG NGHỆ THÔNG TIN

------🙙🕮🙛------



**BÁO CÁO THỰC NGHIỆM**

**MÔN HỌC: PHÂN TÍCH DỮ LIỆU LỚN**

**ĐỀ TÀI:** **DỰ BÁO MỨC LƯƠNG CỦA NGÀNH**

**PHÂN TÍCH DỮ LIỆU BẰNG MÔ HÌNH HỒI QUY TUYẾN TÍNH ĐA BIẾN**

.

**Giảng viên hướng dẫn :** TS. Nguyễn Mạnh Cường

**Lớp :** 20241IT6077002

**Nhóm thực hiện :** Nhóm 04

**Thành viên :** 1. Nguyễn Đắc Trường - 2021605543

2. Nguyễn Thanh Tân - 2021604844

3. Nguyễn Mỹ Hạnh - 2021600647

***Hà Nội - 2024***

# MỤC LỤC

[MỤC LỤC 2](#_Toc186189117)

[DANH MỤC HÌNH ẢNH 4](#_Toc186189118)

[DANH MỤC BẢNG BIỂU 6](#_Toc186189119)

[LỜI CẢM ƠN 7](#_Toc186189120)

[LỜI NÓI ĐẦU 8](#_Toc186189121)

[CHƯƠNG 1: TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU VÀ BÀI TOÁN DỰ BÁO MỨC LƯƠNG 10](#_Toc186189122)

[1.1 Tổng quan về phân tích dữ liệu 10](#_Toc186189123)

[1.1.1 Định nghĩa về phân tích dữ liệu 10](#_Toc186189124)

[1.1.2 Quy trình phân tích dữ liệu 10](#_Toc186189125)

[1.2 Tổng quan về bài toán dự báo mức lương của ngành phân tích dữ liệu 11](#_Toc186189126)

[1.2.1 Giới thiệu chung 11](#_Toc186189127)

[1.2.2. Mục tiêu của bài toán 11](#_Toc186189128)

[1.2.3. Tính thực tiễn của bài toán 12](#_Toc186189129)

[1.3 Phát biểu bài toán 12](#_Toc186189130)

[1.3.1 Xác định đầu vào, đầu ra của bài toán 12](#_Toc186189131)

[1.3.3 Các miền ứng dụng của bài toán 13](#_Toc186189132)

[1.4 Kết quả của bài toán 13](#_Toc186189133)

[CHƯƠNG 2 CÁC PHƯƠNG PHÁP KỸ THUẬT 14](#_Toc186189134)

[2.1 Phương hướng giải quyết bài toán 14](#_Toc186189135)

[2.2 Các phương pháp giải quyết bài toán 15](#_Toc186189136)

[2.2.1 Hồi quy Ridge 15](#_Toc186189137)

[2.2.2. Hồi quy Lasso 16](#_Toc186189138)

[2.2.3 Random Forest 17](#_Toc186189139)

[2.2.3 Hồi quy tuyến tính 19](#_Toc186189140)

[2.3 Lựa chọn mô hình giải quyết bài toán 22](#_Toc186189141)

[CHƯƠNG 3 THỰC NGHIỆM VÀ ĐÁNH GIÁ 24](#_Toc186189142)

[3.1 Giới thiệu quy trình thực hiện 24](#_Toc186189143)

[3.2 Môi trường thực nghiệm 25](#_Toc186189144)

[3.3 Dữ liệu thực nghiệm 26](#_Toc186189145)

[3.4. Thực nghiệm 27](#_Toc186189146)

[3.4.1 Tiền xử lý dữ liệu 27](#_Toc186189147)

[3.4.2 Phân tích mô tả 29](#_Toc186189148)

[3.4.3 Phân tích hồi quy 35](#_Toc186189149)

[3.4.4 Mô hình Random Forest Regression 39](#_Toc186189150)

[3.5 Kết quả thực nghiệm 39](#_Toc186189151)

[CHƯƠNG 4 CHƯƠNG TRÌNH DEMO 42](#_Toc186189152)

[4.1 Giới thiệu thư viện Tkinter trong python 42](#_Toc186189153)

[4.2 Phân tích thiết kế hệ thống 42](#_Toc186189154)

[4.2.1 Biểu đồ use case tổng quát 42](#_Toc186189155)

[4.2.2 Mô tả chi tiết use case 42](#_Toc186189156)

[4.3 Thiết kế giao diện hệ thống 46](#_Toc186189157)

[4.4 Build ra file .exe 47](#_Toc186189158)

[KẾT LUẬN 54](#_Toc186189159)

[TÀI LIỆU THAM KHẢO 55](#_Toc186189160)

# DANH MỤC HÌNH ẢNH

[Hình 1.1 Quy trình phân tích dữ liệu 10](#_Toc186165072)

[Hình 2.1 Hình minh họa nguyên lý hoạt động của Random Forest 18](#_Toc186165073)

[Hình 3.1 Google Colab 25](#_Toc186165074)

[Hình 3.2 Ngôn ngữ Python 25](#_Toc186165075)

[Hình 3.3 minh họa Google Drive 26](#_Toc186165076)

[Hình 3.4 Biểu đồ biểu diễn missing value của bộ dữ liệu với từng cột 27](#_Toc186165077)

[Hình 3.5 Chọn và xử lý cột có dữ liệu kiểu số 28](#_Toc186165078)

[Hình 3.6 điền khuyết giá trị cho cột  work\_year, và salary\_in\_usd 29](#_Toc186165079)

[Hình 3.7 Biểu đồ pie chart thể hiện phân phối tỷ lệ của cột work\_year 29](#_Toc186165080)

[Hình 3.8 Biểu đồ phân phối của cột experience\_level 30](#_Toc186165081)

[Hình 3.9 Hình biểu đồ cột thể hiện tỷ lệ phân bố thời gian làm việc 30](#_Toc186165082)

[Hình 3.10 thể hiện phân phối của tỷ lệ việc làm từ xa 31](#_Toc186165083)

[Hình 3.11 Top 5 công việc phổ biến nhất 31](#_Toc186165084)

[Hình 3.12 Bảng thống kê cho tập dữ liệu 32](#_Toc186165085)

[Hình 3.13 biểu đồ cột thể hiện phân bố chức danh công việc qua các năm 33](#_Toc186165086)

[Hình 3.14 biểu đồ thể hiện xu hướng việc làm qua các năm 34](#_Toc186165087)

[Hình 3.15 biểu đồ thể hiện xu hướng mức lương trung vị qua các năm 34](#_Toc186165088)

[Hình 3.16 Vai trò công việc có mức lương cao nhất 35](#_Toc186165089)

[Hình 3.17 Kiểm tra bản ghi trùng lặp 35](#_Toc186165090)

[Hình 3.18 Xóa các bản ghi trùng lặp 36](#_Toc186165091)

[Hình 3.25 thử trên mô hình Random Forest Regression 39](#_Toc186165092)

[Hình 3.26 Kết quả khi sử dụng mô hình Random Forest 39](#_Toc186165093)

[Hình 4.1 Biểu đồ use tổng quát 42](#_Toc186165094)

[Hình 4.2 Giao diện hệ thống 47](#_Toc186165095)

[Hình 4.3 Giao diện khi Đọc file CSV thành công 48](#_Toc186165096)

[Hình 4.4 Giao diện phân tích mô tả 49](#_Toc186165097)

[Hình 4.5 Giao diện vẽ Histogram 50](#_Toc186165098)

[Hình 4.6 Giao diện vẽ Boxplot 51](#_Toc186165099)

[Hình 4.7 Kết quả dự báo 52](#_Toc186165100)

[Hình 4.8 Giao diện Dự báo 53](#_Toc186165101)

# DANH MỤC BẢNG BIỂU

[Bảng 3.1 Kết quả thực nghiệm 39](#_Toc186165169)

[Bảng 4.1 Bảng mô tả use case Đọc file csv 42](#_Toc186165170)

[Bảng 4.2 Bảng phân tích use case phân tích mô tả 43](#_Toc186165171)

[Bảng 4.3 Bảng phân tích use case vẽ Histogram 44](#_Toc186165172)

[Bảng 4.4 Bảng phân tích use case vẽ Boxplot 44](#_Toc186165173)

[Bảng 4.5 Bảng phân tích use case Clear 45](#_Toc186165174)

[Bảng 4.6 Bảng phân tích use case Dự báo 45](#_Toc186165175)

[Bảng 4.7 Bảng phân tích use case Thoát 46](#_Toc186165176)

# LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn quý thầy, cô Đại học Công Nghiệp Hà Nội, đặc biệt là các giảng viên trong Khoa Công nghệ thông tin, đã tạo điều kiện và hỗ trợ chúng em trong quá trình thực hiện đề tài này.

Chúng em xin bày tỏ lòng biết ơn sâu sắc đối với giảng viên hướng dẫn - TS. Nguyễn Mạnh Cường. Thầy đã tận tình hướng dẫn, cung cấp những kiến thức quý báu và lời khuyên hữu ích, giúp chúng em hoàn thành tốt nhiệm vụ. Bên cạnh đó, chúng em cũng xin cảm ơn các bạn sinh viên trong Khoa Công nghệ thông tin đã đóng góp ý kiến, giúp chúng em thực hiện đề tài đạt hiệu quả hơn.

Thông qua việc thực hiện đề tài "**Dự báo mức lương của ngành phân tích dữ liệu bằng mô hình hồi quy tuyến tính đa biến**", chúng em đã rèn luyện được kỹ năng tư duy phân tích, xử lý dữ liệu và trình bày thông tin một cách logic và rõ ràng. Những kiến thức và kinh nghiệm thu thập từ đề tài này sẽ tiếp tục hỗ trợ chúng em trong tương lai, không chỉ trong học tập mà còn trong sự nghiệp và cuộc sống.

Một lần nữa, chúng em xin trân trọng cảm ơn sự hướng dẫn và định hướng của quý thầy cô và các bạn sinh viên khoa Công nghệ thông tin. Chúng em rất mong nhận được những ý kiến đóng góp để đề tài được hoàn thiện hơn.

Chúng em xin trân trọng cảm ơn!

|  |
| --- |
| **Nhóm thực hiện** |
| *Nhóm 4* |
| *Nguyễn Đắc Trường* |
| *Nguyễn Thanh Tân* |
| *Nguyễn Mỹ Hạnh* |

# LỜI NÓI ĐẦU

Trong bối cảnh nền kinh tế số và công nghệ đang phát triển nhanh chóng, ngành phân tích dữ liệu đã trở thành một trong những lĩnh vực quan trọng với nhu cầu nhân lực ngày càng tăng. Việc hiểu và dự báo mức lương trong ngành phân tích dữ liệu không chỉ giúp các doanh nghiệp và nhà tuyển dụng có cái nhìn tổng quan về thị trường lao động mà còn hỗ trợ các cá nhân định hướng nghề nghiệp phù hợp. Mức lương của các nhà phân tích dữ liệu phụ thuộc vào nhiều yếu tố khác nhau như kinh nghiệm, kỹ năng chuyên môn, trình độ học vấn và địa điểm làm việc.

Trong những năm gần đây, với sự tiến bộ vượt bậc của công nghệ học máy và trí tuệ nhân tạo, các mô hình phân tích và dự báo dựa trên dữ liệu đã mang lại những phương pháp tiếp cận mới, giúp nâng cao độ chính xác và hiệu quả trong việc dự đoán các yếu tố liên quan đến thị trường lao động. Mô hình hồi quy tuyến tính đa biến là một trong những công cụ hữu ích, cho phép phân tích mối quan hệ giữa các biến số và dự báo mức lương trong ngành phân tích dữ liệu một cách khoa học và có căn cứ.

Xuất phát từ nhu cầu thực tiễn này, đề tài "**Dự báo mức lương của ngành phân tích dữ liệu bằng mô hình hồi quy tuyến tính đa biến**" được thực hiện nhằm xây dựng một mô hình dự báo mức lương dựa trên các yếu tố cụ thể trong ngành phân tích dữ liệu. Đề tài không chỉ đóng góp vào việc phát triển các công cụ hỗ trợ trong nghiên cứu thị trường lao động mà còn góp phần thúc đẩy việc ứng dụng các công nghệ tiên tiến trong việc quản lý và phân tích dữ liệu về nhân lực trong thời đại công nghệ số.

Nội dung của quyển báo cáo sẽ bao gồm các chương như sau:

**Chương 1: Tổng quan về đề tài**

Trong chương 1, nhóm tiến hành khái quát nội dung về phân tích dữ liệu và phát biểu bài toán, xác định dữ liệu đầu vào, dữ liệu đầu ra của bài toán, các đặc điểm của bài toán.

**Chương 2: Các phương pháp kỹ thuật**

Sau khi đã phát biểu và xác định rõ ràng được yêu cầu của bài toán, nhóm sẽ trình bày các kỹ thuật và mô hình hồi quy tuyến tính đa biến dùng để phân tích và dự báo.

**Chương 3: Thực nghiệm và đánh giá**

Tại chương này, nhóm tập trung trình bày về quá trình thực nghiệm cũng như các kết quả đạt được với kỹ thuật giải quyết bài toán được đề xuất ở chương 2. Tiến hành so sánh kết quả thực nghiệm thu được từ phương pháp đã được đề xuất giải quyết bài toán.

**Phần kết luận**

Cuối cùng trong phần kết luận, nhóm sẽ tổng hợp các kết quả đạt được, các hướng phát triển và mở rộng đề tài nghiên cứu trong tương lai.

Qua quá trình nghiên cứu và thực hiện đề tài "**Dự báo mức lương của ngành phân tích dữ liệu bằng mô hình hồi quy tuyến tính đa biến**", chúng em đã có cơ hội nâng cao kiến thức về phân tích dữ liệu và ứng dụng các mô hình dự báo trong lĩnh vực nhân sự. Đề tài này không chỉ giúp chúng em hiểu rõ hơn về quá trình thu thập, xử lý và phân tích dữ liệu, mà còn giúp em nhận thức rõ hơn về tiềm năng ứng dụng của các mô hình hồi quy trong việc dự đoán mức lương một cách chính xác và hợp lý. Em hy vọng rằng báo cáo này sẽ cung cấp một công cụ hữu ích cho các nhà tuyển dụng và người lao động, hỗ trợ họ trong việc hoạch định kế hoạch nghề nghiệp và đưa ra các quyết định tài chính phù hợp. Điều này cũng là một bước quan trọng của nhóm trong việc mở ra những hướng nghiên cứu mới, góp phần vào sự phát triển bền vững của ngành nhân sự và phân tích dữ liệu.

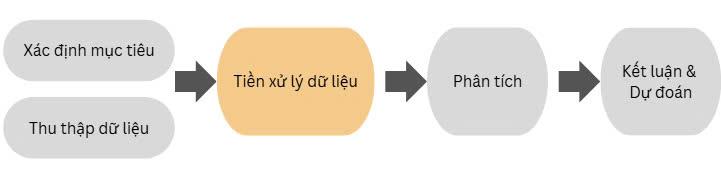
# CHƯƠNG 1: TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU VÀ BÀI TOÁN DỰ BÁO MỨC LƯƠNG

## 1.1 Tổng quan về phân tích dữ liệu

### 1.1.1 Định nghĩa về phân tích dữ liệu

Phân tích dữ liệu chuyển đổi dữ liệu thô thành thông tin chuyên sâu hữu ích. Quá trình này bao gồm đa dạng các công cụ, công nghệ và quy trình được áp dụng để tìm kiếm xu hướng và giải quyết vấn đề bằng dữ liệu. Phân tích dữ liệu có thể định hình các quy trình kinh doanh, cải thiện khả năng ra quyết định và thúc đẩy tăng trưởng doanh nghiệp.[1]

### 1.1.2 Quy trình phân tích dữ liệu

**

Hình 1.1 Quy trình phân tích dữ liệu

**Quy trình phân tích dữ liệu thường gồm các bước sau: [1]**

- **Xác định mục tiêu và vấn đề**: Xác định rõ mục tiêu phân tích và câu hỏi mà phân tích dữ liệu cần giải quyết. Việc này giúp định hướng toàn bộ quy trình và đảm bảo rằng các bước tiếp theo sẽ tập trung vào mục tiêu ban đầu.

- **Thu thập dữ liệu**: Tìm kiếm và thu thập dữ liệu từ các nguồn phù hợp, có thể là từ các cơ sở dữ liệu nội bộ, trang web, khảo sát, hoặc từ các nguồn dữ liệu công khai. Đảm bảo dữ liệu thu thập có liên quan và đáng tin cậy.

- **Làm sạch và tiền xử lý dữ liệu**: Xử lý các lỗi, dữ liệu thiếu, hoặc các giá trị ngoại lệ trong dữ liệu. Bước này cũng bao gồm chuẩn hóa dữ liệu và xử lý các định dạng không đồng nhất để dữ liệu sẵn sàng cho quá trình phân tích.​

- **Phân tích dữ liệu**: Là quá trình chuyển đổi dữ liệu thô thành thông tin chuyên sâu. Có năm loại phân tích chính:

1. Phân tích mô tả: Giúp nắm bắt các sự kiện đã và đang xảy ra, thường thông qua biểu đồ và trực quan hóa dữ liệu.
2. Phân tích chẩn đoán: Tìm hiểu nguyên nhân của sự kiện bằng các kỹ thuật như tìm sâu và khai thác dữ liệu.
3. Phân tích dự đoán: Sử dụng dữ liệu quá khứ để dự báo xu hướng tương lai với các công cụ như máy học và lập mô hình dự đoán.
4. Phân tích theo quy định: Tập trung vào việc đề xuất hành động hoặc giải pháp tối ưu dựa trên dữ liệu đã phân tích. Kỹ thuật này thường sử dụng các mô hình tối ưu hóa và mô phỏng, kết hợp với các phương pháp ra quyết định nhằm giúp tổ chức hoặc cá nhân lựa chọn phương án hành động hiệu quả nhất trong một tập hợp các lựa chọn có sẵn.
5. Phân tích thời gian thực: Là quá trình xử lý và phân tích dữ liệu ngay tại thời điểm dữ liệu được thu thập, nhằm đưa ra các thông tin, cảnh báo hoặc quyết định tức thì. Phương pháp này thường được áp dụng trong các lĩnh vực như tài chính, chăm sóc sức khỏe, quản lý chuỗi cung ứng hoặc các hệ thống IoT, nơi thời gian phản hồi nhanh đóng vai trò quan trọng.

**- Kết luận và dự đoán:** Dựa trên phân tích và thông tin từ dữ liệu, chúng ta có thể rút ra kết luận, hiểu rõ hơn về tình hình, đưa ra dự đoán cho tương lai.

## 1.2 Tổng quan về bài toán dự báo mức lương của ngành phân tích dữ liệu

### 1.2.1 Giới thiệu chung

Bài toán dự báo mức lương là một trong những ứng dụng nổi bật của machine learning và khoa học dữ liệu. Hiện nay, sự phát triển vượt bậc của công nghệ thông tin đã khiến ngành phân tích dữ liệu trở thành một lĩnh vực hấp dẫn với mức cầu lao động cao. Tuy nhiên, mức lương của các chuyên gia trong ngành này chịu ảnh hưởng bởi nhiều yếu tố như kinh nghiệm làm việc, trình độ học vấn, vị trí địa lý, kỹ năng chuyên môn, và yêu cầu cụ thể của từng công ty.

Dự báo mức lương không chỉ giúp các ứng viên đánh giá giá trị bản thân trên thị trường lao động, mà còn hỗ trợ doanh nghiệp xây dựng chính sách lương bổng hợp lý, nhằm thu hút và giữ chân nhân tài. Hơn nữa, bài toán này mang lại giá trị thực tiễn khi kết hợp giữa dữ liệu lớn và các thuật toán học máy để đưa ra những dự báo đáng tin cậy và hiệu quả.

## 1.2.2. Mục tiêu của bài toán

- Xây dựng một mô hình dự báo mức lương chính xác dựa trên các yếu tố quan trọng như kinh nghiệm, học vấn, kỹ năng, khu vực làm việc, và ngành nghề cụ thể.

- Đưa ra các khuyến nghị cụ thể cho doanh nghiệp trong việc thiết kế bảng lương phù hợp với thị trường và nhu cầu tuyển dụng.

- Tạo ra công cụ trực quan hóa dữ liệu (dashboard) hỗ trợ việc phân tích và trình bày kết quả một cách dễ hiểu cho người dùng cuối.

## 1.2.3. Tính thực tiễn của bài toán

- Ngành phân tích dữ liệu đang ngày càng đóng vai trò quan trọng trong mọi lĩnh vực như tài chính, y tế, giáo dục, và sản xuất. Việc áp dụng các công nghệ dự báo mức lương không chỉ giúp doanh nghiệp tối ưu hóa chi phí tuyển dụng mà còn đảm bảo sự hài lòng của nhân viên thông qua các chính sách đãi ngộ minh bạch và hợp lý.

- Ngoài ra, bài toán này còn mang lại giá trị thực tiễn đối với các chuyên gia và sinh viên đang theo đuổi lĩnh vực phân tích dữ liệu. Họ có thể sử dụng các kết quả dự báo để định hướng nghề nghiệp, xây dựng kế hoạch phát triển kỹ năng cá nhân, và xác định mục tiêu lương bổng phù hợp với năng lực.

- Thông qua việc ứng dụng các công cụ và phương pháp phân tích dữ liệu lớn, bài toán này cung cấp một cái nhìn toàn diện về xu hướng lương bổng trong ngành phân tích dữ liệu, góp phần vào sự phát triển bền vững của cả doanh nghiệp và người lao động.

## 1.3 Phát biểu bài toán

### 1.3.1 Xác định đầu vào, đầu ra của bài toán

- Đầu vào là các đặc trưng (features) gồm: work\_year, job\_title, experience\_level, salary\_currency, employee\_residence, remote\_ratio, company\_location

- Đầu ra là mức lương dự báo (salary) của công việc khoa học dữ liệu.

**1.3.2 Khó khăn và thách thức**

- Mối quan hệ phi tuyến tính: Mô hình hồi quy tuyến tính giả định mối quan hệ tuyến tính giữa các đặc trưng và mức lương, nhưng thực tế, mối quan hệ này có thể phi tuyến tính (ví dụ: mức lương không tăng đều theo từng năm kinh nghiệm).

- Dữ liệu không đồng nhất và thiếu dữ liệu: Dữ liệu có thể không đồng nhất về định dạng, và một số đặc trưng như "kỹ năng" hoặc "địa điểm" có thể thiếu hoặc không đầy đủ, gây khó khăn trong việc chuẩn hóa và xử lý.

- Overfitting hoặc Underfitting: Mô hình hồi quy tuyến tính có thể bị overfitting (quá khớp với dữ liệu huấn luyện) hoặc underfitting (không đủ khả năng mô phỏng mối quan hệ phức tạp), ảnh hưởng đến độ chính xác của dự báo.

- Biến động thị trường và yếu tố ngoài dự đoán: Các yếu tố không thể dự đoán, như thay đổi trong nền kinh tế, tình hình tuyển dụng hoặc các sự kiện toàn cầu (ví dụ: đại dịch), có thể gây biến động lớn trong mức lương.

- Vấn đề phân phối không đồng đều: Mức lương có thể phân phối không đồng đều giữa các ngành, vùng địa lý, hoặc công ty, khiến mô hình khó dự báo chính xác cho tất cả các trường hợp.

### 1.3.3 Các miền ứng dụng của bài toán

- Quản lý nhân sự và tuyển dụng: Các công ty có thể sử dụng mô hình dự báo mức lương để xác định mức lương hợp lý cho các vị trí khoa học dữ liệu, từ đó xây dựng chiến lược tuyển dụng và giữ chân nhân viên hiệu quả hơn.

- Phân tích thị trường lao động: Các tổ chức nghiên cứu và cơ quan chính phủ có thể sử dụng mô hình này để phân tích các xu hướng trong ngành khoa học dữ liệu, giúp hiểu rõ hơn về sự thay đổi mức lương theo khu vực, ngành nghề, hoặc năm.

- Đào tạo và phát triển nghề nghiệp: Các trung tâm đào tạo có thể sử dụng dự báo mức lương để giúp học viên và sinh viên hiểu rõ hơn về triển vọng nghề nghiệp và mức lương dự kiến, từ đó đưa ra các khuyến nghị về kỹ năng cần học và lựa chọn nghề nghiệp.

- Tư vấn lương và phúc lợi: Các công ty tư vấn nhân sự có thể áp dụng mô hình này để đưa ra các chiến lược và lời khuyên về lương bổng, phúc lợi cho các tổ chức, giúp họ điều chỉnh mức lương để cạnh tranh trên thị trường lao động.

- Phân tích ngành công nghiệp và dự báo xu hướng: Các tổ chức nghiên cứu ngành công nghiệp có thể sử dụng mô hình để phân tích xu hướng lương trong lĩnh vực khoa học dữ liệu, qua đó giúp các nhà đầu tư và các công ty xác định cơ hội đầu tư và phát triển trong lĩnh vực này.

## 1.4 Kết quả của bài toán

Bài toán dự báo mức lương trong ngành phân tích dữ liệu đã đạt được nhiều kết quả đáng chú ý.Kết quat mức lương dự báo phụ thuộc vào các đặc trưng đầu vào như kinh nghiệm, trình độ học vấn, và vị trí địa lý.

Phân tích dữ liệu cũng chỉ ra rằng các yếu tố ảnh hưởng lớn nhất đến mức lương bao gồm kinh nghiệm làm việc, trình độ học vấn, và vị trí địa lý. Cụ thể, những người có từ 5 năm kinh nghiệm trở lên thường nhận mức lương trung bình cao hơn 20-30% so với những người mới vào nghề.Ngoài ra yếu tố làm Part Time hay Full Time cũng ảnh hưởng nhiều đến mức lương, người làm Full Time luôn có mức lương cao hơn người làm Part Time.

Bên cạnh đó, sự chênh lệch mức lương giữa các ngành nghề phụ trong lĩnh vực phân tích dữ liệu cũng được thể hiện rõ. Chẳng hạn, kỹ sư dữ liệu thường nhận mức lương trung bình cao hơn khoa học dữ liệu khoảng 10%.

Cuối cùng, kết quả từ bài toán cũng cung cấp một khung lương đề xuất cụ thể theo từng mức kinh nghiệm và vị trí địa lý. Những kết quả này không chỉ phản ánh sự thành công của bài toán mà còn có giá trị ứng dụng thực tiễn, giúp các cá nhân và doanh nghiệp tham khảo để đưa ra những quyết định hợp lý trong việc tuyển dụng và đề xuất mức lương.

# CHƯƠNG 2 CÁC PHƯƠNG PHÁP KỸ THUẬT

## 2.1 Phương hướng giải quyết bài toán

Phương hướng giải quyết trong bài toán Dự báo mức lương của ngành khoa học dữ liệu bằng mô hình hồi quy tuyến tính đa biến có thể được thực hiện theo một chuỗi các bước cẩn thận và chi tiết dưới đây:

1. Thu thập và xử lý dữ liệu

- Thu thập dữ liệu từ các nguồn đáng tin cậy như các trang tuyển dụng, khảo sát mức lương, và dữ liệu thống kê ngành nghề.

- Làm sạch dữ liệu, bao gồm xử lý dữ liệu bị thiếu, loại bỏ các giá trị ngoại lai, và chuẩn hóa các biến đầu vào.

- Chuyển đổi dữ liệu thành các định dạng phù hợp để sử dụng trong các mô hình hồi quy tuyến tính, như mã hóa biến phân loại (categorical variables) và chuẩn hóa biến số học (numerical variables).

2. Phân tích mô tả

- Sử dụng các phương pháp thống kê để hiểu rõ hơn về phân phối và mối quan hệ giữa các biến trong dữ liệu.

- Trực quan hóa dữ liệu bằng biểu đồ, đồ thị để xác định các xu hướng, mẫu hình (patterns) và sự bất thường (anomalies).

3. Phân tích dự báo hồi quy

- Xây dựng mô hình hồi quy tuyến tính đa biến với các biến độc lập là các yếu tố ảnh hưởng đến mức lương (như kinh nghiệm, kỹ năng, học vấn) và biến phụ thuộc là mức lương.

- Sử dụng tập dữ liệu huấn luyện (training dataset) để đào tạo mô hình, đồng thời áp dụng các kỹ thuật chọn biến (feature selection) để tối ưu hóa hiệu suất mô hình.

- Kiểm tra mô hình trên tập dữ liệu kiểm định (validation dataset) để đánh giá độ chính xác và khả năng dự báo.

4. Đánh giá và kiểm chứng

- Sử dụng các chỉ số đánh giá mô hình như Mean Squared Error (MSE), R-squared, và Mean Absolute Error (MAE) để kiểm tra hiệu suất của mô hình.

5. Phát triển hệ thống ứng dụng

- Xây dựng giao diện người dùng (UI) để dễ dàng nhập dữ liệu và nhận kết quả dự báo mức lương.

- Tích hợp mô hình dự báo vào hệ thống phần mềm hoặc ứng dụng web, hỗ trợ doanh nghiệp và cá nhân trong việc tra cứu và dự đoán mức lương.

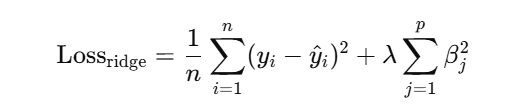
## 2.2 Các phương pháp giải quyết bài toán

### 2.2.1 Hồi quy Ridge

Hồi quy Ridge là một dạng mở rộng của hồi quy tuyến tính, sử dụng L2 regularization để điều chỉnh mô hình và ngăn chặn hiện tượng overfitting. Cụ thể, nó thêm một phần phạt vào hàm mất mát của mô hình, phần phạt này là tổng bình phương của các hệ số hồi quy (β\_j^2) nhân với một hệ số điều chỉnh λ. Mục tiêu chính của phương pháp này là giảm độ phức tạp của mô hình, giúp mô hình học tốt hơn từ dữ liệu và có khả năng tổng quát cao hơn.[4]

Hồi quy Ridge có hai mục tiêu chính: **giảm overfitting** và **xử lý đa cộng tuyến**. Overfitting xảy ra khi mô hình quá khớp với dữ liệu huấn luyện, dẫn đến việc mô hình không thể tổng quát tốt trên dữ liệu mới. Ridge giải quyết vấn đề này bằng cách hạn chế độ lớn của các hệ số hồi quy, làm cho mô hình đơn giản hơn và ít bị ảnh hưởng bởi nhiễu trong dữ liệu. Đối với vấn đề đa cộng tuyến, khi các biến độc lập có mối tương quan mạnh với nhau, phương pháp này giúp giảm phương sai của các ước lượng hệ số, từ đó tăng độ ổn định của mô hình.[4]

**Hàm mất mát**

****

****Sai số dự đoán (Mean Squared Error)

****thành phần phạt, kiểm soát độ lớn của hệ số

Ý nghĩa của λ:

* λ=0: Mô hình trở thành hồi quy tuyến tính bình thường.
* λ lớn: Các hệ số hồi quy bị giảm mạnh về gần 0 (nhưng không bằng 0).

Trong đó, λ là hệ số điều chỉnh, ảnh hưởng đến mức độ mạnh của phần phạt đối với các hệ số. MSE đại diện cho sai số dự đoán giữa giá trị thực tế và giá trị dự đoán của mô hình, trong khi phần phạt giúp kiểm soát độ lớn của các hệ số hồi quy. Nếu λ bằng 0, mô hình trở thành hồi quy tuyến tính thông thường, không có sự điều chỉnh độ lớn của các hệ số. Khi λ lớn, phần phạt L2 trở nên quan trọng hơn, các hệ số hồi quy bị giảm mạnh về gần 0, nhưng không hoàn toàn bằng 0, giúp mô hình tránh việc có các hệ số lớn gây overfitting.[4]

Tóm lại, hồi quy Ridge không chỉ giúp giảm thiểu overfitting mà còn cải thiện tính ổn định của mô hình trong trường hợp có đa cộng tuyến giữa các biến độc lập, qua đó nâng cao khả năng dự báo chính xác trên dữ liệu chưa thấy.

### 2.2.2. Hồi quy Lasso

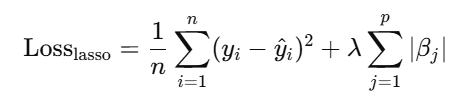
Hồi quy Lasso sử dụng L1 regularization, thêm giá trị tuyệt đối của hệ số hồi quy (βj​∣) vào hàm mất mát.[4]

Mục tiêu của Lasso:[4]

- Chọn lọc biến: Bằng cách đẩy một số hệ số hồi quy về bằng 0.

- Giảm overfitting: Như Ridge, nhưng đồng thời loại bỏ các biến không quan trọng.

**Hàm mất mát**

****

****Sai số dự đoán (Mean Squared Error)

****Thành phần phạt, kiểm soát độ lớn của hệ số

**Ý nghĩa của λ:**

* λ=0: Mô hình trở thành hồi quy tuyến tính bình thường.
* λ lớn: Một số hệ số hồi quy bị đẩy về 0, loại bỏ các biến không quan trọng.

**So sánh Ridge và Lasso**

Bảng 2.1 So sánh Hồi quy Ridge và hồi quy Lasso

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **Hồi quy Ridge** | **Hồi quy Lasso** |
| **Regularization** | L2 | L1 |
| **Đối phó đa cộng tuyến** | Hiệu quả | Kém hơn Ridge |
| **Chọn lọc biến** | Không | Có |
| **Khi nào dùng** | Tất cả các biến đều có giá trị | Nhiều biến không quan trọng |
| **Tác động lên hệ số hồi quy** | Giám độ lớn, không bằng 0 | Một số hệ số bằng 0 |

### 2.2.3 Random Forest

Random Forest là một mô hình học máy thuộc nhóm ensemble learning, sử dụng nhiều cây quyết định (Decision Trees) để thực hiện dự đoán. Nó có thể được sử dụng cho cả bài toán phân loại và hồi quy, nổi tiếng với hiệu suất tốt, khả năng xử lý dữ liệu phức tạp và giảm nguy cơ overfitting.[6]

**1. Nguyên lý hoạt động**

Random Forest được xây dựng dựa trên hai kỹ thuật chính:

- Bagging (Bootstrap Aggregating):

+ Chọn ngẫu nhiên các tập con từ dữ liệu gốc với phép lấy mẫu lặp (with replacement).

+Xây dựng một cây quyết định trên mỗi tập con.

+Tích hợp kết quả từ nhiều cây:

Phân loại: Bằng cách bỏ phiếu đa số (majority vote).

Hồi quy: Bằng cách tính trung bình kết quả.

- Random Feature Selection:

+Tại mỗi nút của cây, chỉ một tập con ngẫu nhiên của các đặc trưng được xem xét để chia tách.

+Điều này giúp giảm sự tương quan giữa các cây và tăng tính đa dạng.[6]

**2. Cách thức hoạt động:**

**Đào tạo mô hình**:

+Tạo nnn tập dữ liệu bootstrap (lấy mẫu ngẫu nhiên từ tập dữ liệu gốc).

+Huấn luyện nnn cây quyết định, mỗi cây trên một tập con khác nhau.

+ Tại mỗi nút phân chia, chọn ngẫu nhiên mmm đặc trưng thay vì sử dụng toàn bộ đặc trưng.

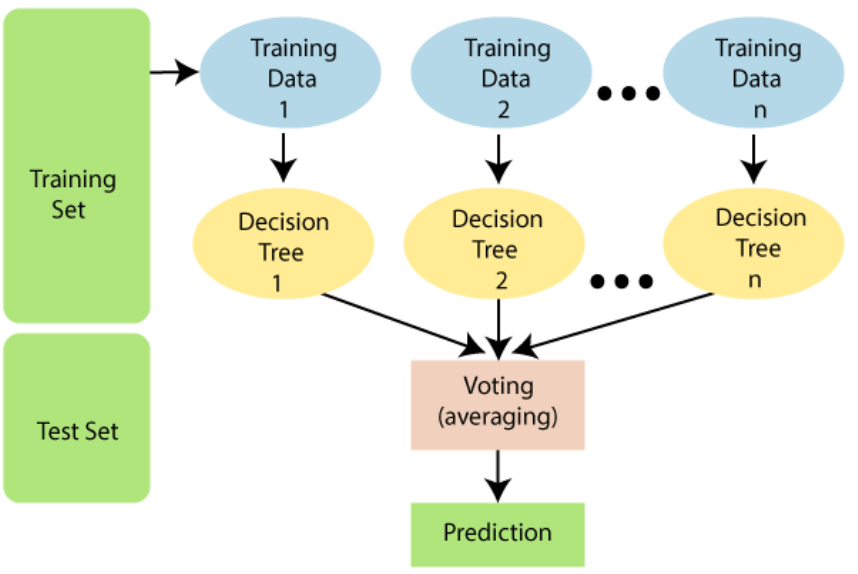
**Dự đoán**:

Với phân loại:

+ Dự đoán lớp từ từng cây và chọn lớp có số phiếu cao nhất.

Với hồi quy:

+ Tính giá trị trung bình từ dự đoán của các cây.



Hình 2.1 Hình minh họa nguyên lý hoạt động của Random Forest

**3. Ưu điểm**

**+** Hiệu suất cao: à một trong những thuật toán tốt nhất để phân loại và hồi quy với dữ liệu phức tạp.

**+** Random Forest giảm overfitting bằng cách trung bình hóa dự đoán từ nhiều cây, thay vì dựa vào một cây duy nhất.

**+** Random Forest có thể xử lý hàng nghìn đặc trưng và dữ liệu lớn mà không cần tiền xử lý phức tạp.

+ Random Forest có thể xác định các đặc trưng quan trọng dựa trên mức độ ảnh hưởng của chúng đến quyết định.

**+** Hoạt động tốt trên cả dữ liệu phân loại và liên tục.

**4. Nhược điểm**

**+** Random Forest yêu cầu nhiều tài nguyên tính toán hơn so với các mô hình đơn giản, đặc biệt với nhiều cây và dữ liệu lớn.

+ Kết quả từ Random Forest khó diễn giải hơn so với một mô hình cây quyết định đơn lẻ.

+ Khi số lượng đặc trưng rất lớn, hiệu quả có thể giảm nếu không tối ưu tham số.

**5. Ứng dụng**

Phân loại:

+ Chẩn đoán y tế (phân loại bệnh từ triệu chứng).

**+** Phân tích khách hàng (dự đoán khách hàng rời bỏ).

Hồi quy:

+ Dự báo giá bất động sản.

**+** Dự đoán doanh thu bán hàng.

Phân tích quan trọng đặc trưng:

+ Xác định các yếu tố quan trọng trong dữ liệu lớn.

### 2.2.3 Hồi quy tuyến tính

Hồi quy tuyến tính" là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Nói cách khác "Hồi quy tuyến tính" là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục. Ví dụ, dự đoán giao thông ở một cửa hàng bán lẻ, dự đoán thời gian người dùng dừng lại một trang nào đó hoặc số trang đã truy cập vào một website nào đó v.v...[5]

**1. Mô hình hồi quy tuyến tính đơn biến**

- Mô hình hóa mối quan hệ giữa một biến độc lập X và một biến phụ thuộc y.

- Công thức: y= +X + ϵ

+ β0​: Hệ số chặn (Intercept), điểm cắt của đường hồi quy với trục y

+ β​: Hệ số hồi quy (Slope), cho biết độ dốc của đường

+ ϵ: Sai số (Residual), đại diện cho sự khác biệt giữa giá trị thực và giá trị dự đoán.

+ y: Biến phụ thuộc (kết quả dự đoán)

+ X: Biến độc lập (kết quả đầu vào)

[5]

- Ưu điểm:

+ Đơn giản, dễ hiểu.

+ Hiệu quả với dữ liệu nhỏ và mối quan hệ tuyến tính rõ ràng.

- Hạn chế:

+ Chỉ phù hợp khi có **một biến độc lập**.

+ Không thể xử lý dữ liệu phức tạp hoặc mối quan hệ phi tuyến.

**2. Mô hình hồi quy tuyến tính đa biến**

- Mô hình hóa mối quan hệ giữa một biến phụ thuộc Y và nhiều biến độc lập X1, X2, …, Xn

- Công thức: Y= + + +…+ +ϵ

+ β​: Hệ số chặn (Intercept), điểm cắt của đường hồi quy với trục y

+ X: Biến độc lập (kết quả đầu vào)

[5]

- Ưu điểm:

+ Mô hình hóa mối quan hệ giữa nhiều biến đầu vào và đầu ra.

+ Hiệu quả với dữ liệu phức tạp hơn

- Nhược điểm:

+ Cần nhiều dữ liệu hơn so với hồi quy đơn biến.

+ Dễ bị ảnh hưởng bởi đa cộng tuyến (multicollinearity) khi các biến độc lập có tương quan cao với nhau.

+ Không phù hợp nếu mối quan hệ giữa biến độc lập và biến phụ thuộc không tuyến tính.

**-** Các giả định của mô hình hồi quy đa biến:[5]

*+* Liên hệ tuyến tính (Linearity): Tồn tại quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc.

+ Phương sai sai số không đổi (Homoscedasticity): Phương sai của phần dư bằng nhau ở tất cả các quan sát.

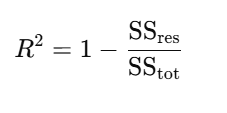
+ Độc lập của sai số (Independence of errors): Phần sai số của các quan sát không có tương quan.

+ Phân phối chuẩn của sai số (Normality): Sai số của các hồi quy phân phối chuẩn.

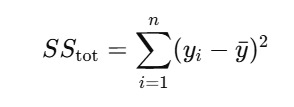
+ Các biến độc lập không có tương quan (Independence of independent variables): biến độc lập không phải là ngẫu nhiên và không có mối quan hệ tuyến tính giữa hai hoặc nhiều biến độc lập.

- Đánh giá hiệu suất mô hình[5]

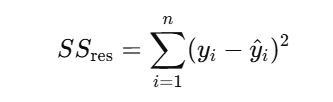
**R-square (R2):** Đo lường tỷ lệ biến thiên của Y được giải thích bởi X



+ Tổng bình phương toàn phần: Đo lường tổng phương sai của dữ liệu thực tế so với giá trị trung bình của Y.



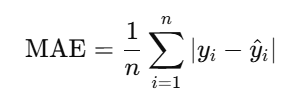
+ Tổng bình phương sai số: đo lường sai lệch giữa giá trị thực và



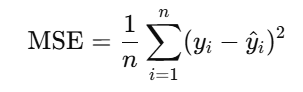
=> Ý nghĩa của R2: nằm trong khoảng từ 0 đến 1:

* **R2 = 1**: Biến phụ thuộc có thể dự đoán mà không có sai số từ biến độc lập
* **R2 = 0**: Biến phụ thuộc không thể dự đoán được biến độc lập
* **R2** (0,1): Vd: R2 = 0.2 nghĩa là 20% phương sai trong y có thể được dự đoán được từ X

**Mean Absolute Error (MAE):** đo lường trung bình của giá trị tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực. [5]



**Mean Squared Error (MSE):** đo lường trung bình bình phương sai số giữa giá trị dự đoán và giá trị thực (nhạy cảm với các giá trị ngoại lai).[5]



**-** Ưu điểm

Hồi quy tuyến tính là một trong những mô hình đơn giản và dễ hiểu nhất trong học máy và thống kê, dưới đây là những ưu điểm chính của hồi quy tuyến tính:

+ Hồi quy tuyến tính rất đơn giản về mặt khái niệm và triển khai, phù hợp với người mới bắt đầu làm quen với phân tích dữ liệu. Có thể dễ dàng cài đặt bằng nhiều ngôn ngữ lập trình như Python, R, hoặc thậm chí bằng tay trong các bảng tính.

+ Hồi quy tuyến tính hoạt động tốt trên tập dữ liệu nhỏ hoặc dữ liệu có mối quan hệ tuyến tính rõ ràng giữa các biến.

+ Có thể dễ dàng mở rộng thành hồi quy đa biến (sử dụng nhiều biến độc lập). Dễ dàng tích hợp với các phương pháp khác (như Polynomial Regression hoặc Regularized Regression: Lasso, Ridge).

+ Huấn luyện mô hình hồi quy tuyến tính có độ phức tạp thấp, thường rất nhanh ngay cả với tập dữ liệu lớn.

- Nhược điểm

Mặc dù hồi quy tuyến tính (Linear Regression) là một phương pháp phổ biến và dễ sử dụng, nhưng nó có một số hạn chế cần lưu ý khi áp dụng vào các bài toán thực tế:

+ Giả định quan hệ tuyến tính: Nếu dữ liệu có mối quan hệ phi tuyến tính, mô hình sẽ không phù hợp và dẫn đến kết quả dự đoán kém.

+ Giả định phương sai không đổi: Hồi quy tuyến tính giả định rằng các sai số có phương sai không đổi. Khi phương sai của sai số thay đổi (heteroscedasticity), mô hình có thể dự đoán không chính xác và làm sai lệch kiểm định thống kê.

+ Nhạy cảm với ngoại lai (Outliers): Các giá trị ngoại lai có thể ảnh hưởng lớn đến các hệ số hồi quy và độ chính xác của mô hình.

+ Đa cộng tuyến (Multicollinearity): Nếu các biến độc lập có mối quan hệ tuyến tính chặt chẽ với nhau, hệ số hồi quy sẽ không ổn định, gây khó khăn trong việc diễn giải mô hình.

## 2.3 Lựa chọn mô hình giải quyết bài toán

Trong quá trình nghiên cứu đề tài "Dự báo mức lương ngành khoa học dữ liệu", nhóm nghiên cứu đã thử nghiệm và đánh giá nhiều mô hình khác nhau, bao gồm Hồi quy tuyến tính (Linear Regression), Ridge, Lasso, và Random Forest. Sau khi xem xét các yếu tố như tính đơn giản, khả năng triển khai nhanh chóng, và hiệu quả tính toán, chúng em đã quyết định chọn mô hình Hồi quy tuyến tính để giải quyết bài toán. Quyết định này được đưa ra sau khi cân nhắc kỹ lưỡng về đặc điểm của bộ dữ liệu và yêu cầu của bài toán.

Hồi quy tuyến tính là một mô hình đơn giản, dễ hiểu và dễ áp dụng. Nó không đòi hỏi kiến thức phức tạp về thuật toán, giúp rút ngắn thời gian triển khai và giảm bớt sự phức tạp trong quá trình giải quyết bài toán. Với đặc thù của bài toán dự báo mức lương ngành khoa học dữ liệu, mô hình Hồi quy tuyến tính có thể cung cấp các kết quả chính xác mà không cần xử lý những yếu tố phức tạp hay yêu cầu tính toán cao như các mô hình phi tuyến.

Bên cạnh đó, hồi quy tuyến tính yêu cầu ít tài nguyên tính toán hơn so với các mô hình phức tạp như Random Forest hay các mô hình đòi hỏi tối ưu hóa siêu tham số như Ridge, Lasso. Thời gian huấn luyện và kiểm tra nhanh chóng của mô hình này đặc biệt hữu ích khi làm việc với các bộ dữ liệu lớn nhưng không quá phức tạp. Vì vậy, việc lựa chọn hồi quy tuyến tính không chỉ giúp tiết kiệm tài nguyên mà còn mang lại hiệu quả cao trong việc dự báo mức lương một cách nhanh chóng và chính xác.

# CHƯƠNG 3 THỰC NGHIỆM VÀ ĐÁNH GIÁ

## 3.1 Giới thiệu quy trình thực hiện

Trong phần này, chúng ta sẽ trình bày chi tiết quy trình thực hiện nghiên cứu từ thu thập dữ liệu đến phân tích kết quả. Quy trình này bao gồm các bước chính sau:

**Bước 1: Thu thập dữ liệu**

Quá trình thu thập dữ liệu là bước đầu tiên và cực kỳ quan trọng trong nghiên cứu. Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau, chẳng hạn như khảo sát, hệ thống thông tin hiện có, các báo cáo tài chính, hoặc cơ sở dữ liệu công khai. Dữ liệu thu thập cần phải đảm bảo tính đầy đủ và chính xác để phục vụ cho các bước phân tích sau này.Dữ liệu được chúng em tìm và thu thập trên KAGGLE.

**Bước 2: Tiền xử lý dữ liệu**

Sau khi thu thập, dữ liệu thường không ở trạng thái sẵn sàng để phân tích. Do đó, tiền xử lý dữ liệu là bước quan trọng để chuẩn bị dữ liệu cho các phân tích tiếp theo. Các bước tiền xử lý thường bao gồm:

+ Làm sạch dữ liệu (loại bỏ các giá trị thiếu, xử lý các giá trị bất thường).

+ Chuyển đổi và chuẩn hóa dữ liệu (đảm bảo các đơn vị, kiểu dữ liệu và phạm vi giá trị đồng nhất).

+ Chuyển đổi dữ liệu thành dạng phù hợp cho các mô hình phân tích.

**Bước 3: Phân tích mô tả**

Phân tích mô tả là bước giúp chúng ta hiểu rõ hơn về đặc điểm của dữ liệu thông qua các thống kê mô tả như trung bình, phương sai, tần suất, và phân phối của các biến. Bước này giúp tạo ra cái nhìn tổng quan về dữ liệu và cung cấp thông tin cơ bản cần thiết để tiến hành các phân tích sâu hơn. Các biểu đồ và bảng thống kê thường được sử dụng trong giai đoạn này để minh họa các kết quả mô tả.

**Bước 4: Phân tích hồi quy**

Phân tích hồi quy là bước quan trọng để xác định mối quan hệ giữa các biến số trong nghiên cứu. Thông qua phân tích hồi quy, chúng ta có thể xây dựng mô hình dự báo và kiểm tra xem các yếu tố nào ảnh hưởng mạnh mẽ nhất đến các kết quả đầu ra. Các kỹ thuật hồi quy phổ biến có thể sử dụng là hồi quy tuyến tính, hồi quy logistic, hoặc hồi quy đa biến, tùy vào mục tiêu nghiên cứu và loại dữ liệu, như trên chúng em lựa chọn mô hình hồi quy tuyến tính.

Quy trình này cung cấp một phương pháp chặt chẽ để phân tích và hiểu dữ liệu, từ đó giúp đưa ra những kết luận chính xác và có cơ sở.

## 3.2 Môi trường thực nghiệm

- Nền tảng thực hiện: Google Colab, một công cụ đám mây mạnh mẽ cho phép chạy mã Python mà không cần cài đặt phần mềm phức tạp. Google Colab cung cấp khả năng sử dụng GPU miễn phí, giúp tăng tốc quá trình huấn luyện mô hình học sâu.



Hình 3.1 Google Colab

- Ngôn ngữ lập trình; Python 3.7 và các thư viện hỗ trợ như NumPy, Pandas, và Matplotlib.

****

Hình 3.2 Ngôn ngữ 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 Minh họa Google Drive

## 3.3 Dữ liệu thực nghiệm

Bộ dữ liệu mà chúng em đang xem xét có tổng cộng 14.838 quan sát và 11 đặc điểm. Trong đó, có 153 chức danh công việc khác nhau, ví dụ như Nhà khoa học dữ liệu (Data Scientist), Nhà phân tích dữ liệu BI (BI Data Analyst), Kỹ sư máy học (ML Engineer), và nhiều hơn nữa. Tập dữ liệu này cũng phân loại các cấp độ kinh nghiệm thành 4 nhóm: MI (Middle), SE (Senior), EN (Entry), và EX (Executive).

Ngoài ra, dữ liệu bao gồm 4 loại hình công việc, gồm FT (Full Time), FL (Freelance), CT (Consultant), và PT (Part Time). Về địa lý, bộ dữ liệu ghi nhận thông tin về 77 địa điểm công ty và 88 quốc gia nơi nhân viên cư trú. Cũng có 3 quy mô công ty: L (Large), M (Medium), và S (Small). Mức lương hàng năm được quy đổi sang USD, dao động từ 15.000 đến 800.000 USD.

*work\_year:* năm của dữ liệu liên quan đến mức lương

*experience\_level:* mức độ kinh nghiệm của nhân viên (mid, senior, expert, ...)

*employment\_type*: loại hình làm việc (part-time, full-time, contract, …)

*job\_title:* chức danh hoặc vai trò của nhân viên trong lĩnh vực khoa học dữ liệu

*salary:* mức lương của nhân viên

*salary\_currency:* đơn vị tiền tệ mà mức lương được biểu thị

*salary\_in\_usd*: mức lương được chuyển sang USD để chuẩn hoá

*employee\_residence:* nơi cư trú của nhân viên

*remote\_ratio:* tỷ lệ làm việc từ xa được phép cho vị trí đó

*company\_location:* địa điểm của công ty

*company\_size:* quy mô của công ty dựa trên số lượng nhân viên hoặc doanh thu

Từ đó ta có thể xác định biến mục tiêu và biến độc lập trong bộ dữ liệu:

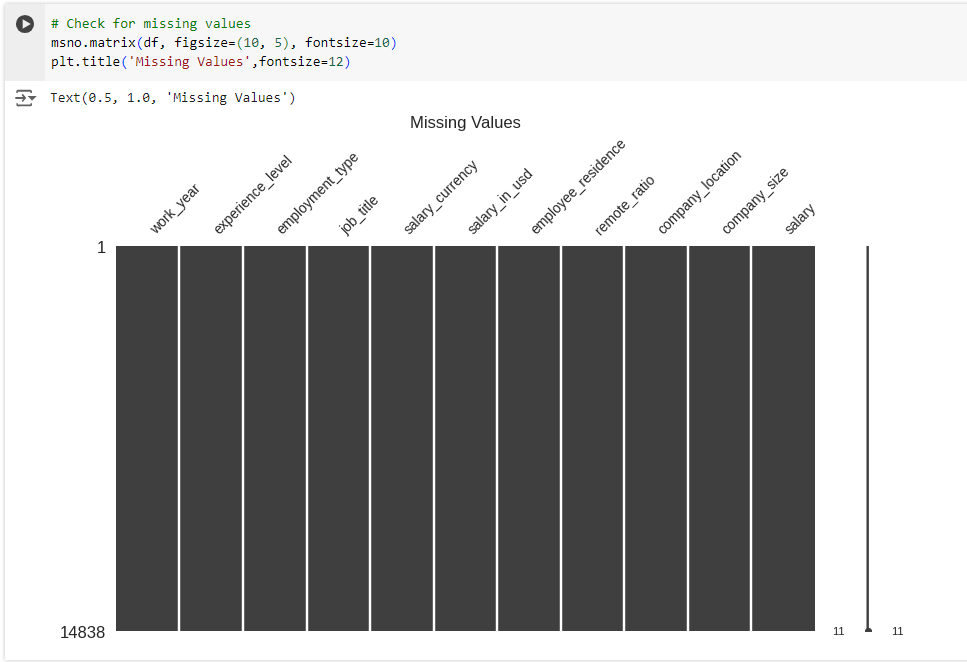
*Biến mục tiêu (Target Variable):* salary\_in\_usd: Đây là biến mục tiêu trong bài toán, đại diện cho mức lương (đã chuẩn hóa sang USD) mà chúng ta cần dự báo.

*Biến độc lập (Independent Variables):* work\_year, experience\_level, employment\_type, job\_title, salary\_currency, employee\_residence, remote\_ratio, company\_location, company\_size.

## 3.4. Thực nghiệm

### 3.4.1 Tiền xử lý dữ liệu

Một điểm đáng chú ý là bộ dữ liệu không có giá trị bị thiếu, với tất cả 14.838 giá trị không rỗng (non-null), điều này đảm bảo rằng quá trình phân tích sẽ không gặp phải vấn đề về dữ liệu thiếu.

**

Hình 3.4 Biểu đồ biểu diễn missing value của bộ dữ liệu với từng cột

Dung lượng bộ nhớ: Tập dữ liệu tiêu tốn **1.2+ MB**, điều này cho thấy nó có kích thước trung bình và dễ dàng xử lý bằng công cụ như Pandas trong Python

**Nhận xét tổng quan:**

+ Chất lượng dữ liệu tốt : Không có giá trị bị thiếu, sẵn sàng để phân tích.

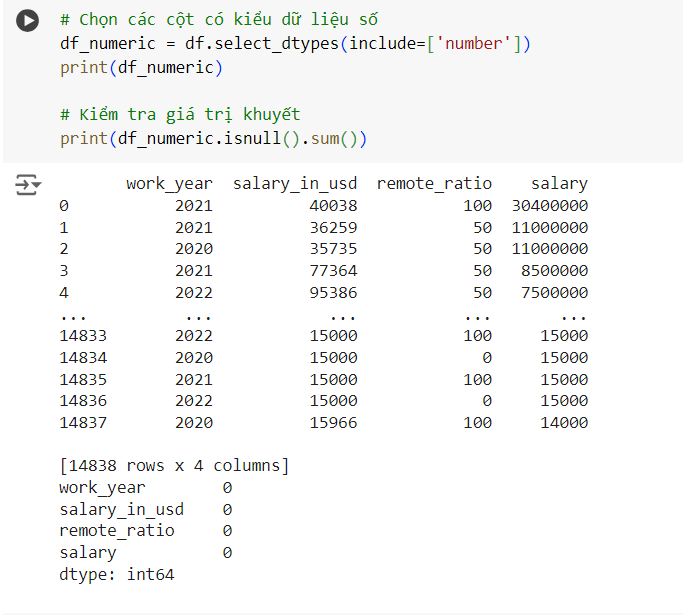
*+* Dữ liệu đa dạng : Bao gồm cả dữ liệu dạng số và chuỗi, phù hợp cho nhiều loại phân tích khác nhau.

+ Đơn vị tiền tệ: Cần kiểm tra mối quan hệ giữa salary\_currency và salary để đảm bảo tính nhất quán.

+ Thông tin hữu ích: Các cột như experience\_level, remote\_ratio, và company\_size có thể giúp phân tích xu hướng thị trường lao động.

**Xử lý dữ liệu liên tục**

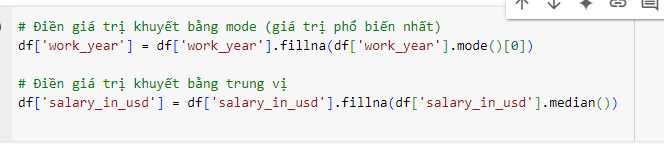
Chúng ta chọn các cột có kiểu dữ liệu số và kiểm tra xem có giá trị khuyết thiếu hay không. Các cột có kiểu số trong bộ dữ liệu bao gồm: work\_year, salary\_in\_usd, remote\_ratio, và salary. Việc kiểm tra giá trị thiếu trong các cột này là cần thiết để đảm bảo tính đầy đủ của dữ liệu.



Hình 3.5 Chọn và xử lý cột có dữ liệu kiểu số

**Phương pháp điền khuyết dữ liệu**:

Với thuộc tính work\_year, chúng tôi sẽ điền khuyết bằng giá trị phổ biến (mode), trong khi thuộc tính salary\_in\_usd sẽ được điền khuyết bằng giá trị trung vị (median) để đảm bảo tính chính xác trong các phân tích sau.



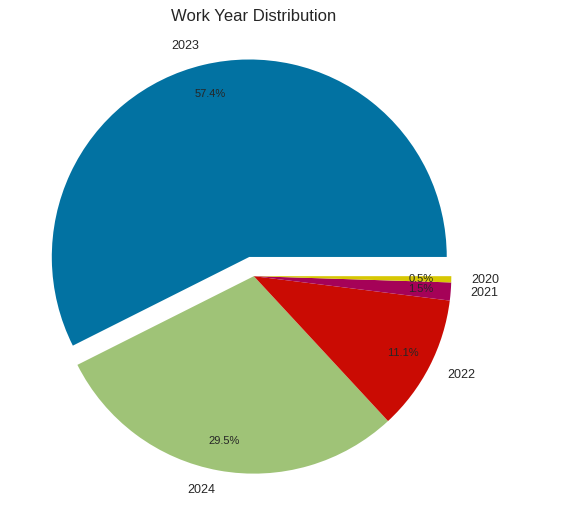
Hình 3.6 Điền khuyết giá trị cho cột  work\_year, và salary\_in\_usd

### 3.4.2 Phân tích mô tả

**3.3.2.1 Phân tích đơn biến**

**Phân phối của cột work\_year**

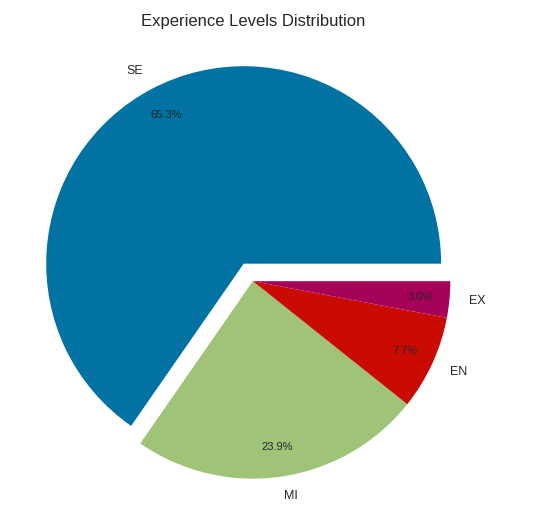
Biểu đồ pie chart thể hiện phân phối tỷ lệ phần trăm của cột work\_year cho thấy rằng chủ yếu dữ liệu liên quan đến năm 2023. Số lượng tham gia khảo sát vào năm 2021 và 2022 rất ít, điều này làm cho bộ dữ liệu chủ yếu phản ánh xu hướng nghề nghiệp của năm 2023 và 2024.

****

Hình 3.7 Biểu đồ pie chart thể hiện phân phối tỷ lệ của cột work\_year

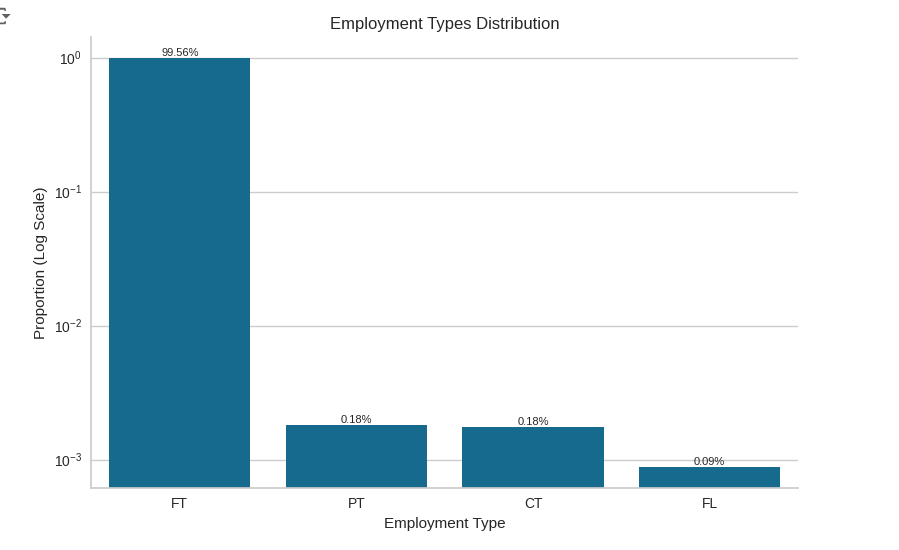
**Phân phối của cột experience\_level**

Biểu đồ pie chart thể hiện tỷ lệ phân phối của các cấp độ kinh nghiệm cho thấy 90% số người tham gia khảo sát là các chuyên gia cấp cao và trung cấp.

****

Hình 3.8 Biểu đồ phân phối của cột experience\_level

**Phân phối của loại hình công việc**

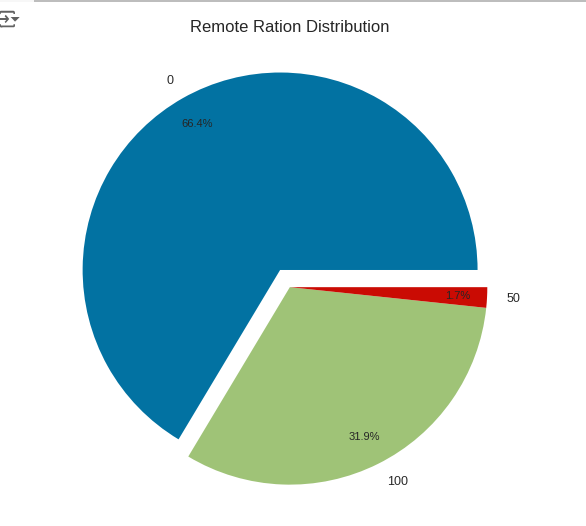
Biểu đồ cột với thang đo logarit sẽ thể hiện tỷ lệ phân bố của các loại hình công việc ****

Hình 3.9 Hình biểu đồ cột thể hiện tỷ lệ phân bố thời gian làm việc

Gần như tất cả mọi người tham gia khảo sát, gần 100%, đều báo cáo làm việc toàn thời gian.

**Phân phối của tỷ lệ làm việc từ xa (remote\_ratio)**

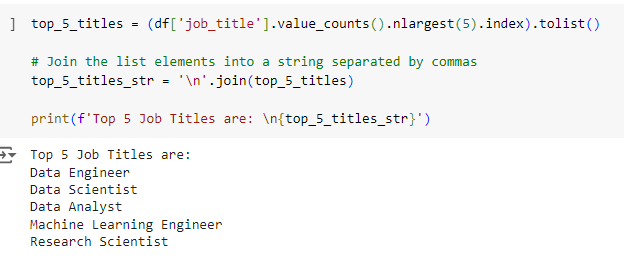
Biểu đồ pie chart thể hiện tỷ lệ làm việc từ xa cho thấy khoảng 66% người tham gia khảo sát cho biết họ làm việc hoàn toàn tại văn phòng, với tỷ lệ làm việc từ xa là tương đối thấp.



Hình 3.10 Thể hiện phân phối của tỷ lệ việc làm từ xa

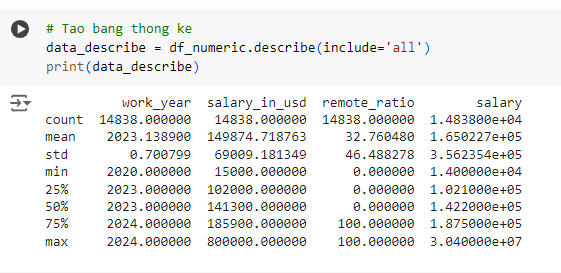
**Top 5 công việc phổ biến nhất**

Dựa trên phân tích của chúng tôi, một trong những mục tiêu là xác định 5 công việc phổ biến nhất trong bộ dữ liệu. Biểu đồ dưới đây hiển thị sự phân bố của các công việc này, với tần suất xuất hiện của mỗi chức danh. Các công việc phổ biến nhất, bao gồm Data Engineer, Data Scientist, và Data Analyst, chiếm tỷ lệ lớn trong bộ dữ liệu.



Hình 3.11 Top 5 công việc phổ biến nhất

**Tạo bảng thống kê các thuộc tính kiểu số:**



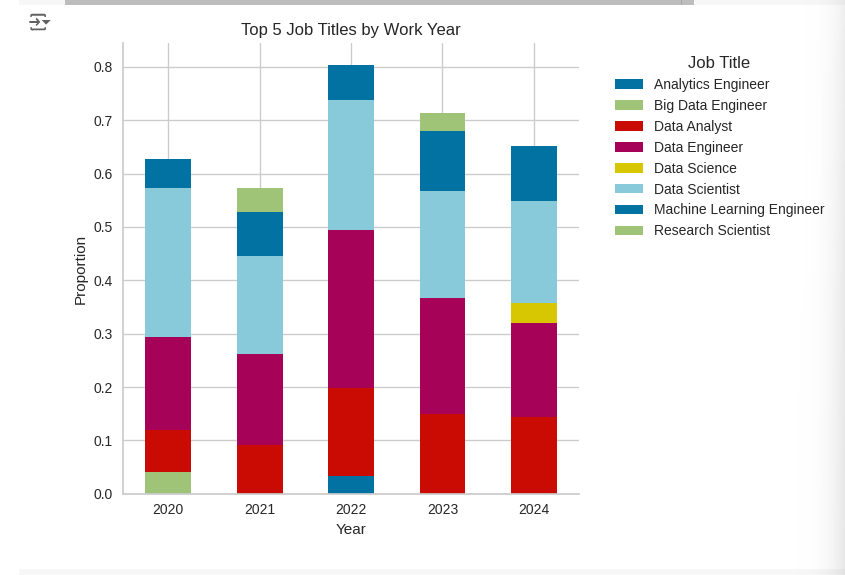
Hình 3.12 Bảng thống kê cho tập dữ liệu

**3.3.2.2. Phân tích đa biến**

**Khám phá Xu hướng và Phân bố Công Việc Qua Thời Gian**

*Phân bố chức danh công việc qua các năm*

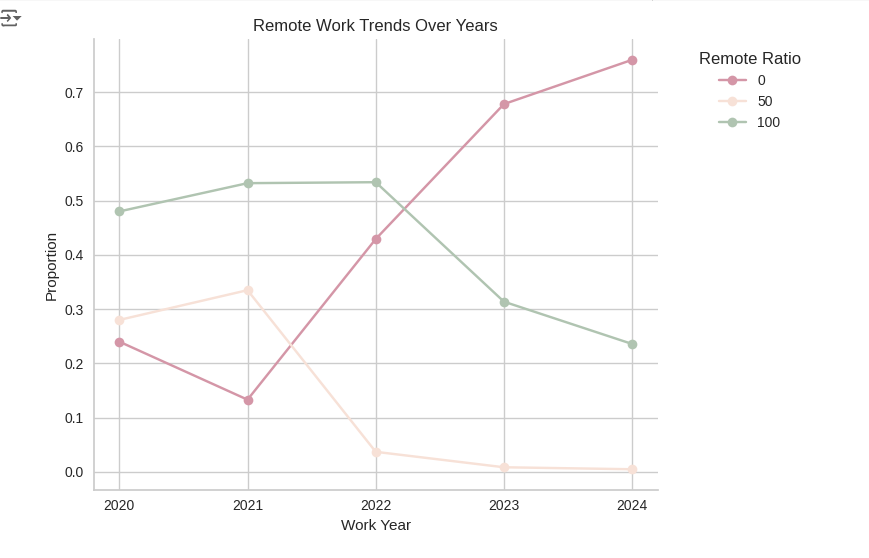
Biểu đồ cột dưới đây sẽ thể hiện sự phân bố của các công việc theo từng năm. Mục tiêu là phân tích các công việc hàng đầu mỗi năm và nhận diện những công việc mới nổi cũng như các công việc giảm dần trong thị trường lao động. Các công việc như Data Scientist và AI Specialist có thể được nhận diện là các vai trò mới nổi trong vài năm qua.

****

Hình 3.13 Biểu đồ cột thể hiện phân bố chức danh công việc qua các năm

**Xu hướng làm việc từ xa qua các năm**

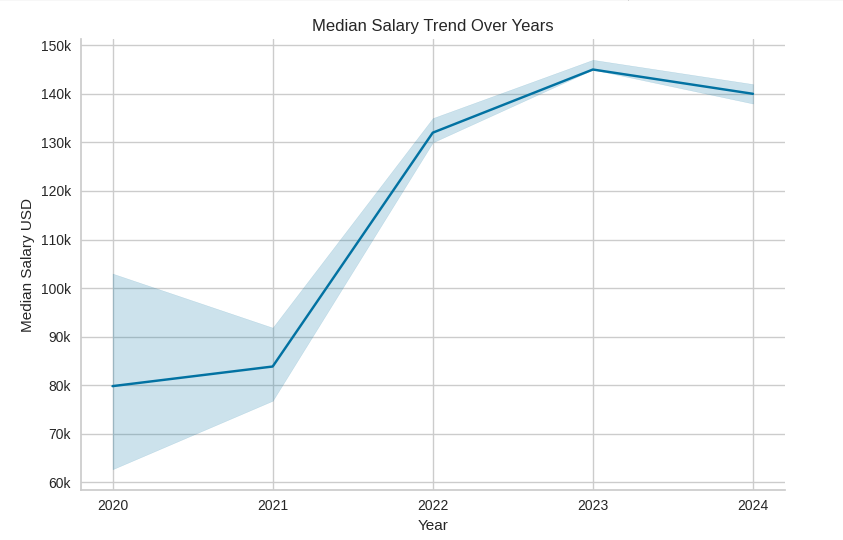
Biểu đồ thể hiện xu hướng làm việc từ xa qua các năm cho thấy sự thay đổi rõ rệt trong cách thức làm việc. Trong giai đoạn 2020-2022, tỷ lệ nhân viên làm việc hoàn toàn từ xa hoặc bán thời gian từ xa tăng lên đáng kể, phản ánh ảnh hưởng của đại dịch đối với hình thức làm việc tại văn phòng. Tuy nhiên, từ năm 2022, xu hướng này bắt đầu giảm, khi nhiều công ty yêu cầu nhân viên quay lại làm việc tại văn phòng, thể hiện sự dịch chuyển trở lại các chuẩn mực làm việc trước đại dịch.



Hình 3.14 Biểu đồ thể hiện xu hướng việc làm qua các năm

**Xu hướng mức lương trung vị qua các năm**

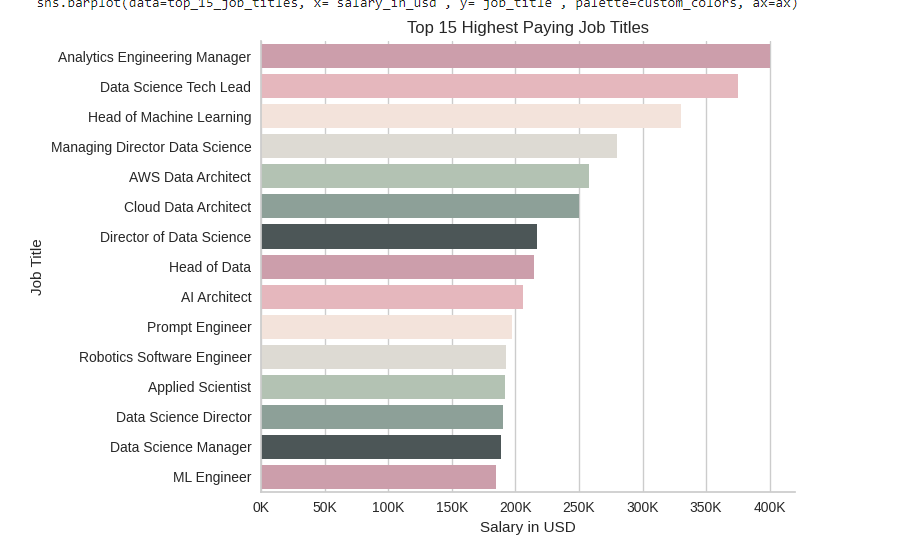
Biểu đồ dưới đây thể hiện xu hướng thay đổi mức lương trung vị qua các năm. Mức lương trung vị đã tăng mạnh từ 80.000 USD vào năm 2020 lên 145.000 USD vào năm 2023, phản ánh sự gia tăng nhu cầu về các chuyên gia trong các lĩnh vực công nghệ và dữ liệu. Tuy nhiên, năm 2024 có dấu hiệu giảm nhẹ, có thể phản ánh sự điều chỉnh của thị trường lao động.

****

Hình 3.15 Biểu đồ thể hiện xu hướng mức lương trung vị qua các năm

**Các vai trò có công việc có mức lương cao nhất**

Phân tích này sẽ nêu bật 15 vị trí có mức lương cao nhất dựa trên báo cáo lương của người tham gia khảo sát.

****

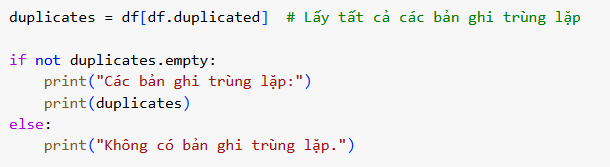
Hình 3.16 Vai trò công việc có mức lương cao nhất

Cuối cùng, phân tích các vị trí công việc có mức lương cao nhất trong bộ dữ liệu cho thấy các vai trò quản lý cấp cao và các chuyên gia kỹ thuật cao, như Kiến trúc sư Dữ liệu, AI, AWS, và Cloud, chiếm ưu thế trong danh sách các công việc có mức lương cao. Điều này phản ánh giá trị cao của các kỹ năng thiết kế hệ thống và cơ sở hạ tầng trong ngành công nghệ.

Thông qua các phân tích trên, chúng ta có thể thấy rõ sự thay đổi trong cấu trúc công việc, mức lương và xu hướng làm việc từ xa qua các năm, đồng thời có cái nhìn chi tiết hơn về các yếu tố ảnh hưởng đến thị trường lao động hiện nay.

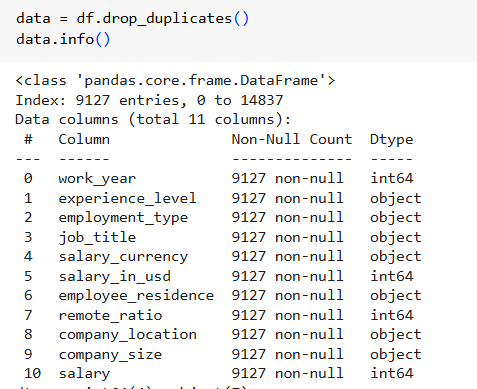
### 3.4.3 Phân tích hồi quy

- Kiểm tra bản ghi trùng lặp



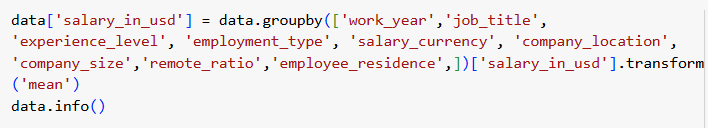
Hình 3.17 Kiểm tra bản ghi trùng lặp

- Xoá các bản ghi trùng lặp

****

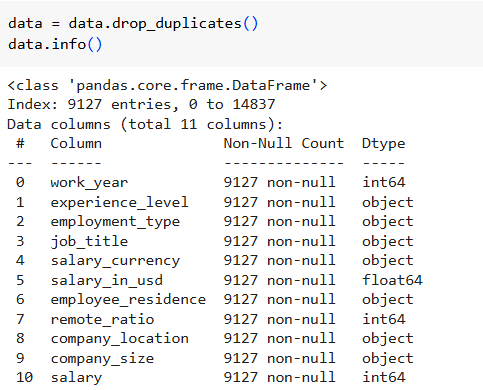
Hình 3.18 Xóa các bản ghi trùng lặp

- Nhóm các bản ghi có cùng giá trị ở các cột thuộc tính để gán trung bình

****

*Hình: Nhóm các bản ghi có cùng giá trị ở các cột thuộc tính*

- Xoá các bản ghi trùng lặp



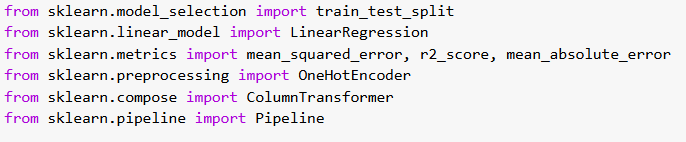
*Hình: Dữ liệu sau khi xóa các bản ghi trùng lặp*

- Xử lý các biến phân loại

****

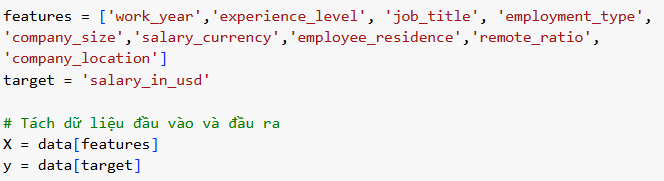
*Hình 3.19 Xử lý các biến phân loại*

- Gọi các thư viện cần thiết

****

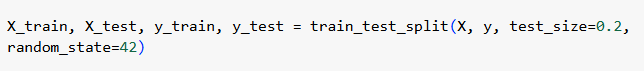
*Hình 3.20 Thư viện cần thiết*

- Tách dữ liệu đầu ra và đầu vào

****

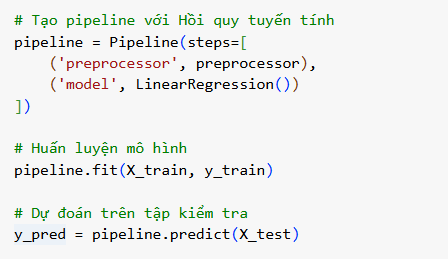
*Hình 3.21 tách dữ liệu đầu vào*

- Chia bộ dữ liệu thành tập train và test với tỉ lệ 8-2

****

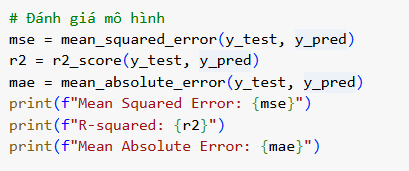
*Hình 3.22 chia tập dữ liệu thành tập train và test*

- Tạo pipeline, huấn luyện mô hình và dự đoán trên tập kiểm tra

****

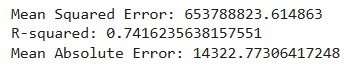
*Hình 3.23 Huấn luyện mô hình*

- Đánh giá mô hình dựa trên các chỉ số

****

*Hình 3.24 Đánh giá mô hình*

**Kết quả hiệu suất mô hình sau khi chạy chương trình:**

****

**Dựa vào kết quả trên ta có thể nhận xét như sau:**

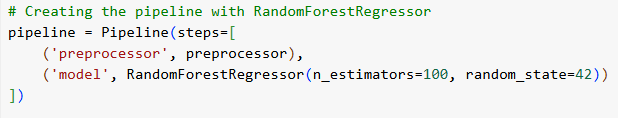
- Mean Squared Error: 653788823 đại diện cho sai số bình phương trung bình, cho biết mức độ khác biệt trung bình giữa giá trị thực tế và giá trị dự đoán của mô hình. Mặc dù giá trị này lớn (do đơn vị là USD và mức lương có thể dao động cao) nhưng bản thân MSE không trực quan bằng các chỉ số khác như MAE. Điều này xảy ra do MSE nhạy cảm với các ngoại lệ (outliers)

- R²: 0.7416 (~74.16%), R² thể hiện tỷ lệ biến thiên của biến mục tiêu (salary\_in\_usd) được giải thích bởi các biến độc lập trong mô hình. Với giá trị 74.16%, mô hình hồi quy tuyến tính đã giải thích được phần lớn sự biến đổi trong mức lương, điều này cho thấy mô hình có hiệu quả tương đối tốt. Tuy nhiên, vẫn còn khoảng 25.84% sự biến đổi không được giải thích, có thể là do các yếu tố không nằm trong tập dữ liệu hoặc các mối quan hệ phi tuyến tính mà mô hình hồi quy tuyến tính không nắm bắt được.

- MAE: 14,322.77 USD, MAE là sai số trung bình tuyệt đối, đo lường sai lệch trung bình giữa giá trị thực và giá trị dự đoán theo cùng đơn vị (USD). Kết quả này cho thấy, trung bình, mức lương dự đoán của mô hình chênh lệch khoảng 14,322.77 USD so với thực tế. Giá trị MAE cao như vậy có thể do dữ liệu có sự đa dạng cao về mức lương.

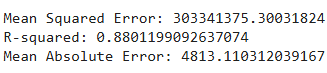
### 3.4.4 Mô hình Random Forest Regression

Import mô hình và thay đổi mô hình Random Forest:



Hình 3.25 Thử trên mô hình Random Forest Regression

Kết quả :



Hình 3.26 Kết quả khi sử dụng mô hình Random Forest

Mô hình Random Forest Regressor đã cho thấy hiệu quả vượt trội hơn hẳn. Hệ số R-squared của mô hình này đạt 0.8801, tức là nó giải thích được khoảng 88.01% biến thiên trong dữ liệu. Đồng thời, cả MSE và MAE của Random Forest Regressor đều thấp hơn rất nhiều so với hồi quy tuyến tính, cho thấy độ chính xác cao hơn và sai số dự đoán thấp hơn.

## 3.5 Kết quả thực nghiệm

Tiến hành thực hiện chạy chương trình từ đầu 10 lần với mỗi mô hình để có thể khách quan hơn trong việc đánh giá.

Bảng 3.1 Kết quả thực nghiệm

|  |  |  |
| --- | --- | --- |
| Lần chạy | Hồi quy tuyến tính | Random Forest Regression |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| 8 |  |  |
| 9 |  |  |
| 10 |  |  |
| Trung bình | R-squared : 0.742 | R-squared : 0.88 |

Mô hình hồi quy tuyến tính, với sự đơn giản và khả năng giải thích cao, vẫn là một lựa chọn rất hữu ích trong nhiều trường hợp. Kết quả của mô hình này cho thấy nó có thể giải thích được phần lớn biến thiên trong dữ liệu và cung cấp một nền tảng tốt để hiểu mối quan hệ giữa các biến. Mặc dù mô hình Random Forest Regressor có hiệu suất cao hơn, nhưng sự phức tạp và yêu cầu tính toán của nó có thể không phù hợp với tất cả các bài toán hoặc môi trường ứng dụng.

Trong tổng kết, hồi quy tuyến tính, với tất cả những ưu điểm của nó, là một công cụ mạnh mẽ và hiệu quả cho các bài toán dự đoán và phân tích dữ liệu. Nó không chỉ mang lại kết quả đáng tin cậy mà còn giúp người dùng dễ dàng hiểu và áp dụng.

# CHƯƠNG 4 CHƯƠNG TRÌNH DEMO

## 4.1 Giới thiệu thư viện Tkinter trong python

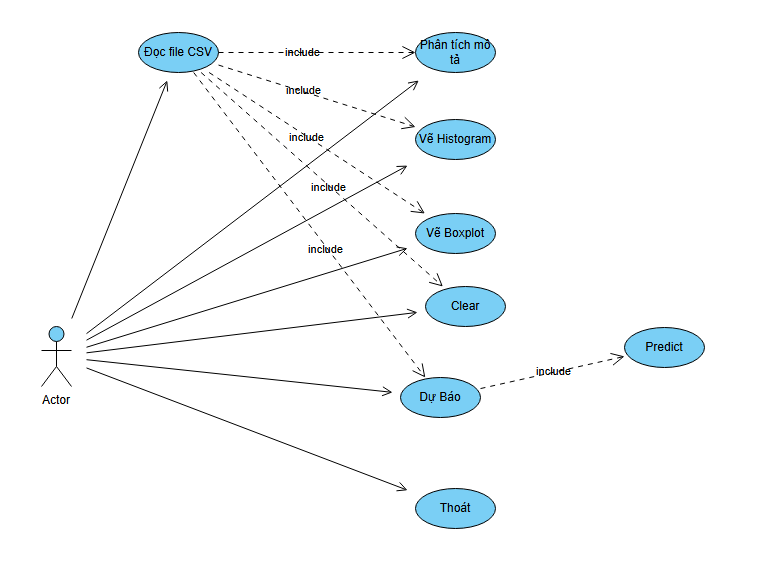
Tkinter là một thư viện trong ngôn ngữ lập trình Python được sử dụng để tạo giao diện đồ họa người dùng (GUI). "Tkinter" là viết tắt của "Tk interface,", một toolkit đồ họa cung cấp các công cụ để phát triển giao diện người dùng.

Tkinter là một phần của thư viện tiêu chuẩn của Python và đã được tích hợp sẵn trong hầu hết các cài đặt Python. Điều này giúp cho Tkinter trở thành một lựa chọn phổ biến cho việc phát triển ứng dụng với giao diện đồ họa đơn giản trong Python.

Một số đặc điểm của Tkinter bao gồm khả năng tạo các thành phần giao diện như cửa sổ, nút, ô văn bản, và các widget khác để tương tác với người dùng. Tkinter cung cấp cả các sự kiện và phương thức để xử lý tương tác người dùng và thay đổi trạng thái của ứng dụng.

## 4.2 Phân tích thiết kế hệ thống

### 4.2.1 Biểu đồ use case tổng quát



Hình 4.1 Biểu đồ use tổng quát

### 4.2.2 Mô tả chi tiết use case

- Mô tả use case Đọc file csv

Bảng 4.1 Bảng mô tả use case Đọc file csv

|  |  |
| --- | --- |
| **Mã use case** | UC1. |
| **Tên use case** | Đọc file CSV |
| **Tóm tắt** | Use case này cho phép lựa chọn file csv để tìm hiểu cũng như dự báo. |
| **Actor** | Người dùng. |
| **Tiền điều kiện** | Ứng dụng phải khởi động. |
| **Đảm bảo tối thiểu** | Nếu người dùng không chọn file thì không thực hiện được các chức năng khác. |
| **Đảm bảo thành công** | Người dùng chọn được 1 file csv. |
| **Kích hoạt** | Tự động. |
| **Luồng sự kiện** | 1. Người dùng click Đọc file Csv.  2. Chọn file .csv rồi chọn Open. |

-Mô tả use case Phân tích mô tả :

Bảng 4.2 Bảng phân tích use case phân tích mô tả

|  |  |
| --- | --- |
| **Mã use case** | UC2 |
| **Tên use case** | Phân tích mô tả |
| **Tóm tắt** | Use case này cho phép người dùng mô tả các thông số cơ bản của bộ dữ liệu. |
| **Actor** | Người dùng |
| **Luồng sự kiện chính** | Người dùng click vào Phân tích mô tả |
| **Điều kiện tiên quyết** | Người dùng phải đọc 1 file csv trước đó |
| **Kết quả đạt được** | Bảng mô tả thống cơ các thông số cơ bản của dữ liệu. |

-Mô tả use case Vẽ histogram:

Bảng 4.3 Bảng phân tích use case vẽ Histogram

|  |  |
| --- | --- |
| **Mã use case** | UC3 |
| **Tên use case** | vẽ Histogram. |
| **Tóm tắt** | Use case này thực hiện vẽ biểu đồ histogram của bộ dữ liệu |
| **Actor** | Người dùng. |
| **Luồng sự kiện chính** | Người dùng click vào vẽ histogram. |
| **Điều kiện tiên quyết** | Người dùng phải đọc 1 file csv trước đó. |
| **Kết quả đạt được** | Các biểu đồ histogram của từng thuộc tinh của bộ dữ liệu. |

- Mô tả use case vẽ Boxplot

Bảng 4.4 Bảng phân tích use case vẽ Boxplot

|  |  |
| --- | --- |
| **Mã use case** | UC4 |
| **Tên use case** | vẽ Boxplot |
| **Tóm tắt** | Use case này thực hiện vẽ biểu đồ Boxlot của bộ dữ liệu |
| **Actor** | Người dùng. |
| **Luồng sự kiện chính** | Người dùng click vào vẽ Boxlot. |
| **Điều kiện tiên quyết** | Người dùng phải đọc 1 file csv trước đó. |
| **Kết quả đạt được** | Các biểu đồ Boxplot của từng thuộc tinh của bộ dữ liệu. |

- Mô tả use case Clear

Bảng 4.5 Bảng phân tích use case Clear

|  |  |
| --- | --- |
| **Mã use case** | UC5 |
| **Tên use case** | Clear |
| **Tóm tắt** | Use case này thực hiện xóa dữ liêu vừa mở, làm mới lại từ đầu |
| **Actor** | Người dùng. |
| **Luồng sự kiện chính** | Người dùng click vào Clear. |
| **Điều kiện tiên quyết** | Người dùng phải đọc 1 file csv trước đó. |
| **Kết quả đạt được** | Không có bất cứ dữ liệu nào trên màn hình cả. |

- Mô tả use case Dự báo

Bảng 4.6 Bảng phân tích use case Dự báo

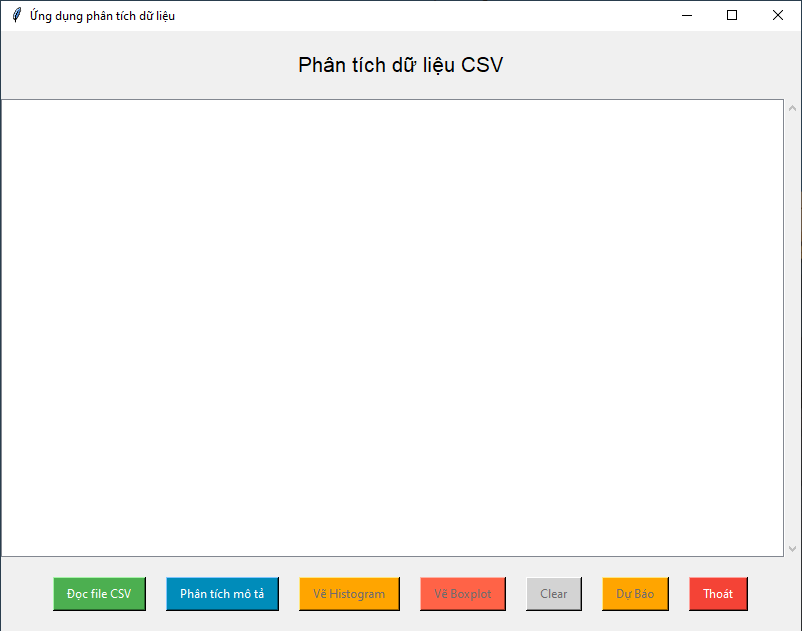
|  |  |
| --- | --- |
| **Mã use case** | UC6 |
| **Tên use case** | Dự báo |
| **Tóm tắt** | Use case này thực hiện dự báo mức lương |
| **Actor** | Người dùng. |
| **Luồng sự kiện chính** | 1.Người dùng click vào Dự báo  2.Người dùng điền thông tin lên from  3.Người dùng click Predict. |
| **Điều kiện tiên quyết** | Người dùng phải đọc 1 file csv trước đó. |
| **Kết quả đạt được** | Mức lương dự báo. |

- Mô tả use case Thoát

Bảng 4.7 Bảng phân tích use case Thoát

|  |  |
| --- | --- |
| **Mã use case** | UC7 |
| **Tên use case** | Thoát. |
| **Tóm tắt** | Use case này thực hiện thoát ứng dụng. |
| **Actor** | Người dùng. |
| **Luồng sự kiện chính** | Người dùng click vào Thoát. |
| **Điều kiện tiên quyết** | Không có điều kiện tiên quyết. |
| **Kết quả đạt được** | Ứng dụng tắt hẳn. |

## 4.3 Thiết kế giao diện hệ thống



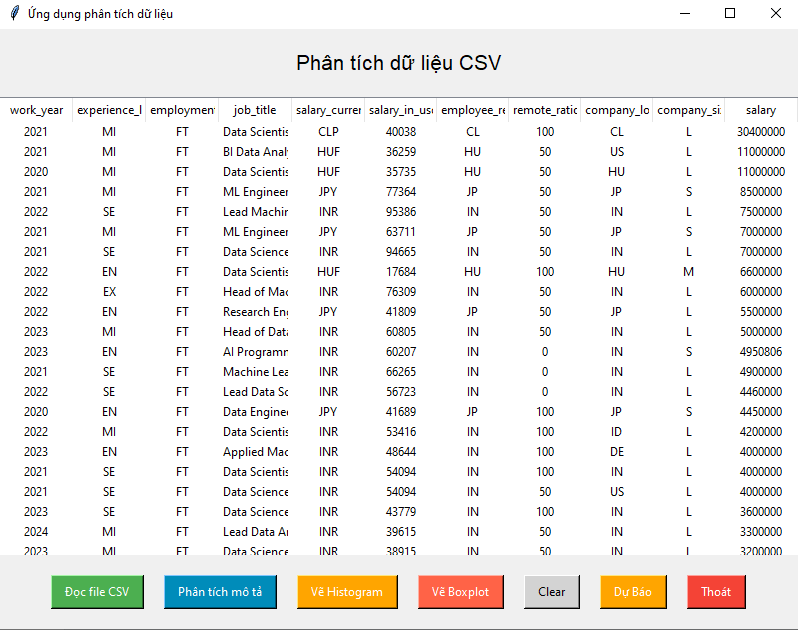
Hình 4.2 Giao diện hệ thống

## 4.4 Build ra file .exe

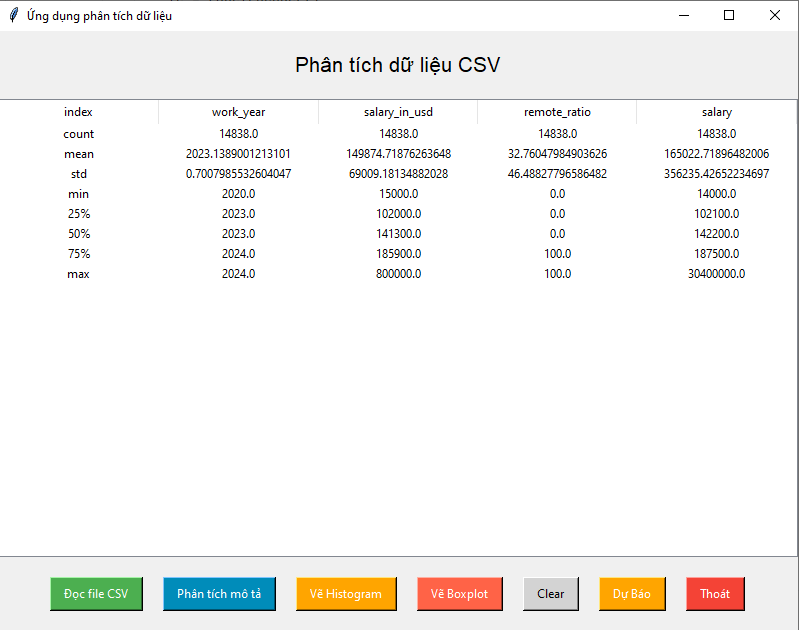
Dowload thư viện “pyinstaller” bằng lệnh : pip install pyinstaller

Rồi thực hiện lệnh sau để build ra .exe

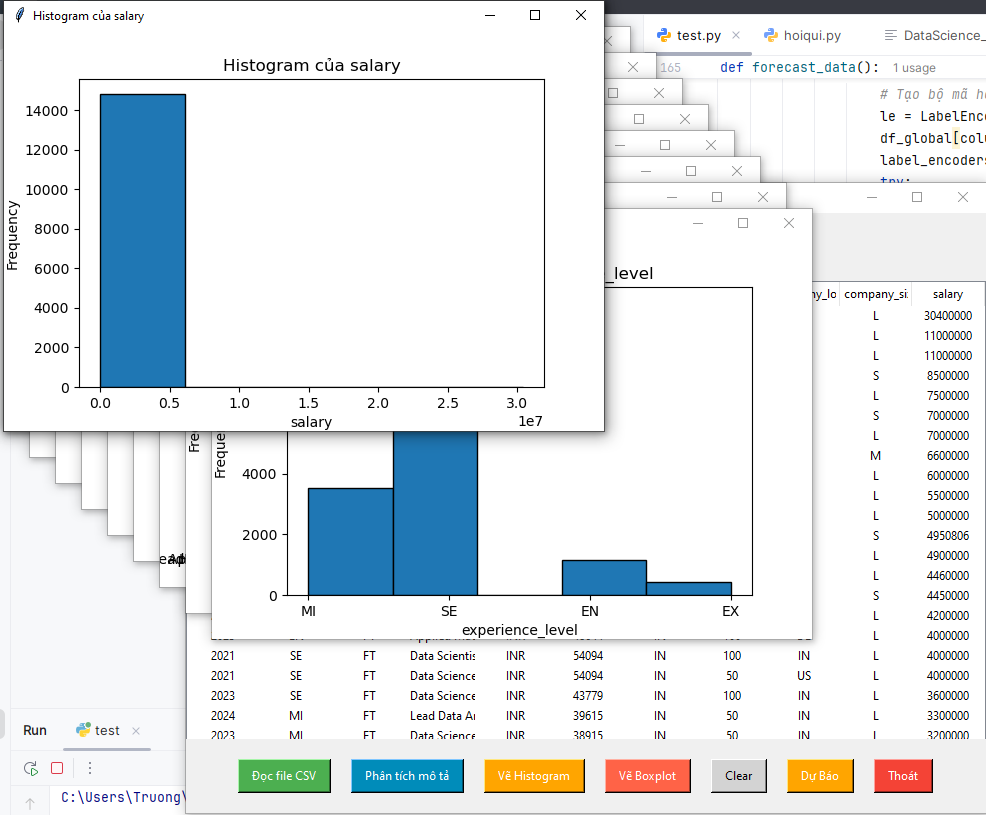




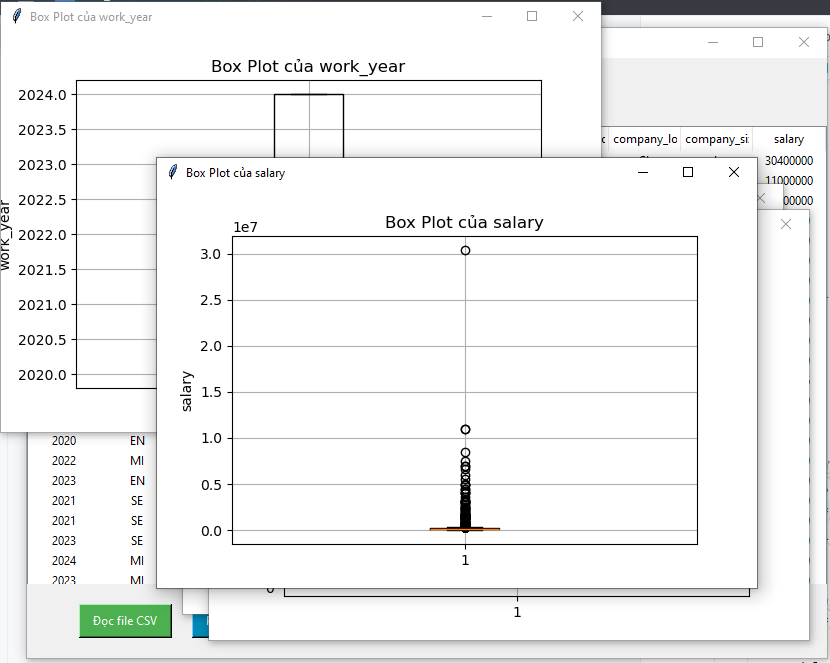
Hình 4.3 Giao diện khi Đọc file CSV thành công



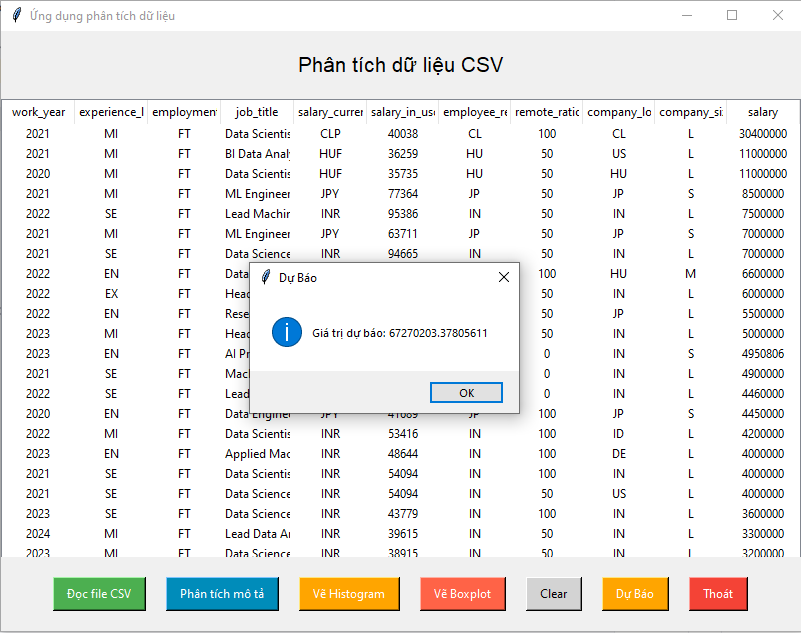
Hình 4.4 Giao diện phân tích mô tả



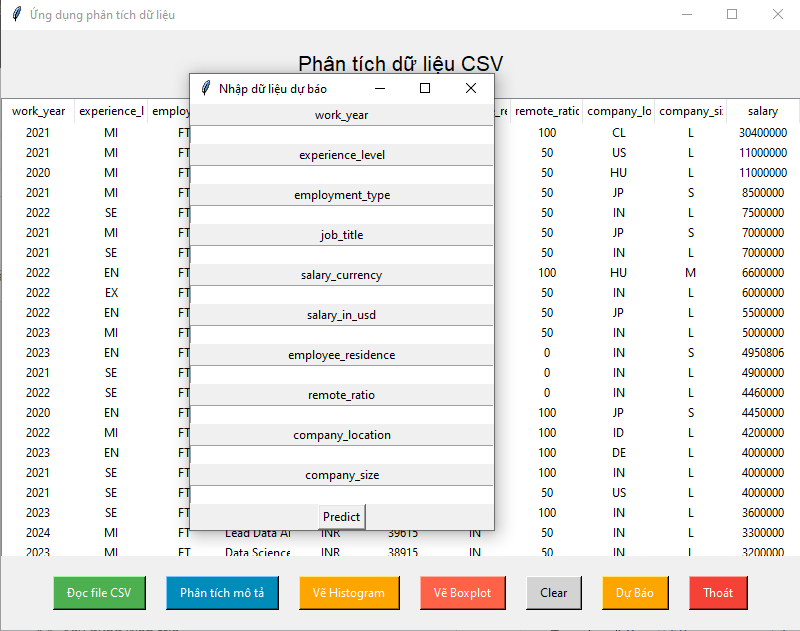
Hình 4.5 Giao diện vẽ Histogram



Hình 4.6 Giao diện vẽ Boxplot



Hình 4.7 Kết quả dự báo



Hình 4.8 Giao diện Dự báo

# KẾT LUẬN

Trong quá trình thực hiện đề tài "Dự báo mức lương sử dụng hồi quy tuyến tính", chúng em đã tập trung nghiên cứu và so sánh hiệu quả của các phương pháp khác nhau trong việc dự đoán mức lương dựa trên các yếu tố đầu vào như kinh nghiệm, trình độ học vấn, và kỹ năng chuyên môn. Mục tiêu của chúng em là phát triển một hệ thống có khả năng dự báo mức lương một cách chính xác và hiệu quả, qua đó hỗ trợ người lao động và nhà tuyển dụng trong việc đưa ra quyết định về mức lương phù hợp.

Qua các thí nghiệm, chúng em đã thu được những kết quả quan trọng. Mô hình hồi quy tuyến tính chứng tỏ là lựa chọn tối ưu với độ chính xác cao, khả năng dự báo tốt và đơn giản trong triển khai. Các chỉ số đánh giá như MSE (Mean Squared Error) và R² (R-squared) đều cho thấy mô hình này có khả năng dự đoán mức lương một cách chính xác và hiệu quả. Điều này khẳng định rằng hồi quy tuyến tính là một phương pháp mạnh mẽ trong việc giải quyết bài toán dự báo mức lương.

Tuy nhiên, trong quá trình thực hiện, chúng em nhận ra rằng mô hình hồi quy tuyến tính chưa đạt hiệu suất cao như kỳ vọng trong một số trường hợp, đặc biệt khi dữ liệu có nhiều yếu tố phi tuyến. Mô hình này có khả năng tổng quát hóa kém hơn trong những tình huống phức tạp, làm giảm độ chính xác của dự báo.

Dựa trên những kết quả và hạn chế đã gặp phải, trong tương lai, chúng em sẽ tập trung vào việc cải thiện chất lượng dữ liệu đầu vào, đồng thời nghiên cứu và tích hợp các kỹ thuật học máy khác để tăng cường khả năng dự báo. Đặc biệt, chúng em sẽ tiếp tục phát triển và tối ưu hóa các mô hình phức tạp hơn để nâng cao hiệu quả dự báo mức lương, giúp hệ thống có thể hoạt động tốt hơn trong các điều kiện thực tế đa dạng và phức tạp.

# TÀI LIỆU THAM KHẢO

**TRANG WEB**

|  |  |
| --- | --- |
| [1] | Website tổng quan về phân tích dữ liệu. URL: <https://mindx.edu.vn/blog/phan-tich-du-lieu-la-gi-quy-trinh-va-vi-du-thuc-te-ve-phan-tich-du-lieu> . (Lần truy cập gần nhất : 28/11/2024). |
| [2] | Website dạy trí tuệ nhân tạo. URL: <https://trituenhantao.io/kien-thuc/decision-tree/> . (Lần truy cập gần nhất : 28/11/2024). |
| [3] | Website giới thiệu về thư viện Tkinter. URL:  <https://www.icantech.vn/kham-pha/tkinter>. (Lần truy cập gần nhất : 29/11/2024). |
| [4] | Website giới thiệu về hồi quy Ridge và hồi quy Lasso. URL: <https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_RidgedRegression.html> (Lần truy cập gần nhất : 01/12/2024). |
| [5] | Website giới thiệu về hồi quy tuyến tính. URL:  <https://aws.amazon.com/vi/what-is/linear-regression/>. (Lần truy cập gần nhất : 02/12/2024). |
| [6] | Website tìm hiểu về thuật toán RadomForest. URL: <https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_RandomForest.html> (Lần truy cập gần nhất : 28/11/2024). |