**Thẻ 1**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**BỘ NÔNG NGHIỆP VÀ PHÁT TRIỂN NÔNG THÔN**

**TRƯỜNG ĐẠI HỌC THỦY LỢI**



BÁO CÁO BÀI TẬP LỚN HỌC PHẦN

DỮ LIỆU LỚN

**ỨNG DỤNG MACHINE LEARNING**

**DỰ ĐOÁN LƯƠNG DỰA TRÊN SỐ NĂM KINH NGHIỆM**

**Nhóm sinh viên thực hiện :** Nhóm 4

ĐỖ HUYỀN CHINH MSV : 2251172257

VƯƠNG TẤT CHIẾN MSV : 2251172256

HOÀNG THÂN VŨ MSV : 2251172560

**Giáo viên hướng dẫn :** TRẦN ANH ĐẠT

Hà nội, 10- 2024

**MỤC LỤC**

[**LỜI NÓI ĐẦU** 3](#_gjdgxs)

[PHẦN 1 : MỞ ĐẦU 4](#_4erguegoqgix)

[**1. Lý do chọn đề tài** . 4](#_l0hw6pngmns)

[**2. Mục tiêu nghiên cứu** 5](#_wqpr6sytqt5z)

[**3. Đối tượng và phạm vi nghiên cứu** 5](#_cb8sar4oayvt)

[PHẦN 2: CƠ SỞ LÝ THUYẾT 6](#_hl6o0nicpx6)

[**1. Lý thuyết học máy** 6](#_7uy0zovict6w)

[**2. Phân loại học máy** 7](#_21fpydrgt6ls)

[**3. Các thuật toán áp dụng cho đề tài** 8](#_ysdql5dj5eri)

[3.1 Hồi quy tuyến tính 8](#_7j7szos5b4hq)

[3.2 Lasso 9](#_zeosa9hpyg6i)

[3.3 Neural Network 10](#_cqfl6zp29xrh)

[3.4 Stacking 11](#_toq5kzv20nyj)

[**4. Các phương pháp đánh giá mô hình học máy cho đề tài** 12](#_az0dw3jufosa)

[4.1 MAE ( Mean Absolute Error) : Trung bình lỗi tuyệt đối 12](#_dx8l3kdvrsel)

[4.2 MSE (Mean Squared Error) - Trung bình bình phương sai số 12](#_9a9co3dcphdd)

[4.3 R2 ( R-squared) : Hệ số xác định 13](#_uwy1ygapuea)

[CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM: 14](#_bmkr0n1qvjls)

[**3.1. Dữ liệu** 14](#_4tdn02ib3tbf)

[3.1.2 Độ tin cậy của dữ liệu 14](#_4q73y8qijlw7)

[3.1.3. Các thuộc tính được thu thập 14](#_sx680yg91eo)

[3.1.4. Tiền xử lí dữ liệu 14](#_gvsyx0tauifq)

[3.1.5. Chia dữ liệu 14](#_splzfymhaqt)

[**3.2. Kết quả thực nghiệm với Linear Regression** 15](#_6228ubpwr45z)

[**3.3. Kết quả thực nghiệm với Lasso Regression** 15](#_d8xw0tvaqocy)

[**3.4. Kết quả thực nghiệm với Neural Network** 16](#_un3q2p3h91d6)

[**3.5. Kết quả thực nghiệm với Stacking** 16](#_3wfoowcuu2xz)

[PHẦN 4: KẾT LUẬN VÀ ĐÁNH GIÁ 17](#_jmszxgkvpn0r)

[**1. Đánh giá** 17](#_e4hzf774y2r1)

[**2. Kết luận** 17](#_ph9qquq88sve)

[2.1 Kết quả đạt được: 17](#_z6lmand9nr52)

[2.2 Những hạn chế: 18](#_x0lezs2xkwso)

[TÀI LIỆU THAM KHẢO](#_6ctf24mvcvh) 19

# LỜI NÓI ĐẦU

Trong bối cảnh phát triển mạnh mẽ của trí tuệ nhân tạo và các ứng dụng học máy, việc dự đoán mức lương dựa trên kinh nghiệm làm việc của nhân sự là một chủ đề thu hút sự quan tâm lớn từ cả các nhà nghiên cứu và doanh nghiệp. Một mô hình dự đoán chính xác không chỉ hỗ trợ các cá nhân có cái nhìn rõ hơn về con đường sự nghiệp của họ mà còn giúp các doanh nghiệp đưa ra các quyết định tuyển dụng và chính sách đãi ngộ hợp lý hơn.

Đề tài "Dự đoán lương dựa trên số năm kinh nghiệm" sẽ tập trung vào việc ứng dụng các phương pháp học máy trong việc phân tích mối quan hệ giữa số năm kinh nghiệm và mức lương của nhân sự trong các lĩnh vực khác nhau. Trong nghiên cứu này, chúng em sẽ sử dụng các phương pháp như hồi quy tuyến tính,Lasso, Neural Network, Stacking. Đồng thời, chúng em cũng sẽ đánh giá hiệu quả của các mô hình dựa trên các chỉ số như R2 (R-squared) , MAE (Mean Absolute Error) ,MSE (Mean Squared Error) .

Thông qua bài tập lớn này, chúng em hy vọng sẽ hiểu sâu hơn về quy trình xây dựng mô hình học máy, cũng như cách xử lý và phân tích dữ liệu thực tế. Đồng thời, kết quả của bài tập cũng có thể cung cấp những thông tin hữu ích cho các tổ chức và cá nhân trong việc đánh giá và dự đoán lương cho các vị trí nhân sự khác nhau.

# PHẦN 1 : MỞ ĐẦU

## Lý do chọn đề tài .

Trong thời đại công nghệ phát triển nhanh chóng, nhu cầu phân tích dữ liệu và dự đoán xu hướng ngày càng trở nên quan trọng đối với các doanh nghiệp và cá nhân. Đặc biệt, việc dự đoán mức lương dựa trên số năm kinh nghiệm có vai trò thiết thực trong lĩnh vực quản trị nhân sự, vì nó không chỉ giúp cá nhân có cơ sở tham khảo để định hướng sự nghiệp mà còn giúp doanh nghiệp thiết lập các chính sách lương thưởng hợp lý.

Chúng em chọn đề tài “Dự đoán lương dựa trên số năm kinh nghiệm” vì những lý do chính sau đây:

1. **Tính ứng dụng cao trong thực tế**: Việc dự đoán mức lương dựa trên kinh nghiệm là nhu cầu phổ biến, có ý nghĩa lớn đối với cả cá nhân trong việc định hướng sự nghiệp và doanh nghiệp trong chiến lược tuyển dụng, đãi ngộ.
2. **Là một bài toán học máy phổ biến và cơ bản**: Đây là một bài toán hồi quy điển hình, giúp chúng em có cơ hội ứng dụng các thuật toán học máy như hồi quy tuyến tính và các kỹ thuật tối ưu khác, từ đó nâng cao kiến thức và kỹ năng thực hành trong lĩnh vực học máy.
3. **Nguồn dữ liệu phong phú**: Đề tài này có thể tận dụng nguồn dữ liệu phong phú từ các nền tảng nghề nghiệp, mạng lưới xã hội và báo cáo nhân sự, từ đó tăng khả năng tiếp cận các dữ liệu thực tế để áp dụng mô hình.
4. **Khả năng mở rộng và phát triển đề tài**: Sau khi hoàn thành mô hình dự đoán cơ bản, chúng em có thể mở rộng đề tài bằng cách tích hợp thêm các yếu tố khác như ngành nghề, vị trí công việc, địa điểm làm việc và trình độ học vấn, từ đó giúp mô hình dự đoán chính xác và sát với thực tế hơn.

Với những lý do trên, chúng em tin rằng việc thực hiện đề tài này không chỉ có ý nghĩa về mặt học thuật, mà còn mang lại những giá trị thiết thực khi ứng dụng vào thực tế, góp phần vào quá trình nghiên cứu và phát triển các mô hình học máy trong lĩnh vực quản lý và phát triển nhân sự

## Mục tiêu nghiên cứu

***2.1 Mục tiêu tổng quát***

Xây dựng mô hình học máy hiệu quả để dự đoán lương thông qua việc nhập số năm kinh nghiệm

***2.2 Mục tiêu cụ thể***

* Tiền xử lý dữ liệu, chuẩn hoá dữ liệu giúp cho mô hình hiệu học quả hơn.
* Áp dụng các phương pháp học máy như : Hồi quy tuyến tính, Lasso, Neural Network , Stacking xây dựng mô hình học máy để dự đoán lương dựa trên số năm kinh nghiệm
* Đánh giá độ chính xác của các mô hình từ đó chọn ra mô hình, tham số tốt nhất để dự đoán lương .

## Đối tượng và phạm vi nghiên cứu

***3.1 Đối tượng nghiên cứu***

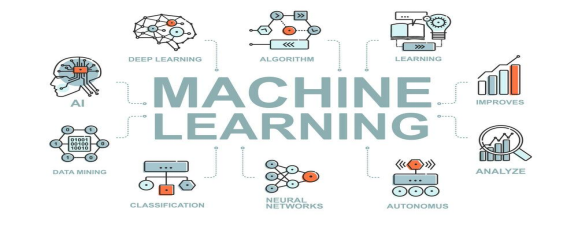
Tìm hiểu , nghiên cứu về mức lương .

***3.2 Phạm vi nghiên cứu***

* Phạm vi đề tài : Nghiên cứu tập trung vào các đối tượng đã đi làm
* Phạm vi thời gian : Nghiên cứu được thực hiện từ 9/2024 đến 10/2024

# PHẦN 2: CƠ SỞ LÝ THUYẾT

## Lý thuyết học máy



Artificial Intelligence (AI) – Trí tuệ nhân tạo là một ngành thuộc lĩnh vực khoa học máy tính, là trí tuệ do con người tạo ra nhằm xử lý các bài toán khoa học và các vấn đề của con người. Điều này có nghĩa là máy tính có thể nhận dạng được hình ảnh thực tế, ngôn ngữ tự nhiên hay mô phỏng hành động trong thế giới thực.

Machine Learning là một nhánh của AI và khoa học máy tính tập trung vào việc sử dụng dữ liệu và thuật toán nhằm mô phỏng cách con người học tập, cải thiện độ chính xác.

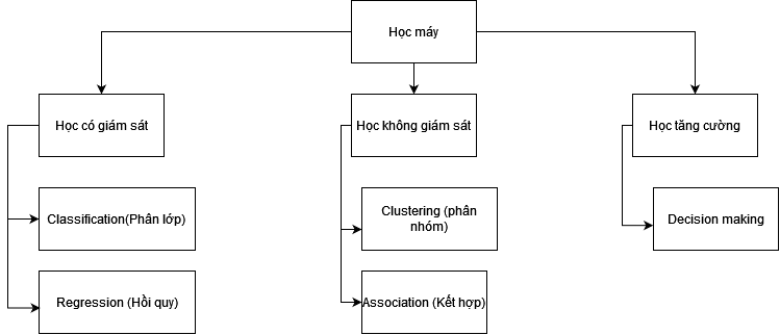
Machine Learning là việc nghiên cứu các thuật toán mà có thể phát triển tự động thông qua kinh nghiệm. Machine learning khám phá các nghiên cứu và xây dựng thuật toán mà có thể học tập từ dữ liệu và tạo ra các dự đoán trên dữ liệu. Dựa trên càng nhiều dữ liệu, Machine learning có thể đưa ra những phản hồi hiệu quả hơn thích ứng hơn và có thể mở rộng hơn.

Ứng dụng của Machine Learning rất phong phú và đa dạng trong các lĩnh vực phân loại, các ứng dụng mạng xã hội, nhận diện hình ảnh, nhận diện giọng nói. Machine Learning giúp phân tích các phép đo của một đối tượng để xác định xem nó thuộc về một danh mục nào để thiết lập mối quan hệ giúp các nhà phân tích sử dụng dữ liệu.

Machine Learning đang ngày càng được sử dụng rộng rãi trong nhiều lĩnh vực vì khả năng xử lý các dữ liệu lớn và phức tạp, đẩy mạnh nỗ lực của chúng ta trong việc giải quyết các vấn đề phức tạp trong đời sống.

## Phân loại học máy

Trong Machine learning có rất nhiều thuật toán và được chia ra làm ba nhóm thuật toán chính là học không giám sát, học có giám sát, bán giám sát và học tăng cường. Dưới đây là hình ảnh tổng quan về các thuật toán:



*Các nhóm thuật toán trong Machine learning*

Sau đây là một số thuật toán phổ biến trong Machine learning:

- Hồi quy tuyến tính (Linear Regression)

- Hồi quy logistic (Logistic Regression)

- Phân cụm K-means (K-means Clustering)

- SVM (Support Vector Machine)

- Cây quyết định (Decision Tree)

Mỗi thuật toán có một cơ chế hoạt động khác nhau tùy thuộc vào mục đích sử dụng. Tuy nhiên, các thuật toán này đều được sử dụng để giải quyết các vấn đề phức tạp và nắm giữ vai trò quan trọng trong các ứng dụng của machine learning

## Các thuật toán áp dụng cho đề tài

### *3.1* Hồi quy tuyến tính

* Là một thuật toán học có giám sát, được sử dụng để giải quyết bài toán dự đoán giá trị của một biến liên tục dựa trên các biến đầu vào (Biến độc lập)
* Mục tiêu của hồi quy tuyến tính là tìm một đường thẳng mô tả quan hệ tuyến tính giữa các biến đầu vào và đầu ra.
* Công thức :



Trong đó :

+ y là giá trị dự đoán

+ x1,x2,....xn là các biến đầu vào độc lập

+ w1,w2,.....wn là các hệ số tương ứng của từng biến đầu vào

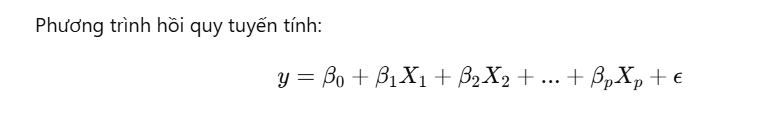
+ b là hằng số tự do

=> Mục tiêu: tìm ra các hệ số wi sao cho sai số giữa giá trị dự đoán và giá trị thực nhỏ nhất

* Ưu điểm
* Là một thuật toán đơn giản, dễ hiểu, tính toán nhanh và dễ mở rộng
* Hiệu quả với các mối quan hệ tuyến tính
* Các hệ số hồi quy có thể giúp dễ dàng diễn giải mối quan hệ giữa từng biến độc lập và biến phụ thuộc, thể hiện tác động của mỗi yếu tố trong mô hình.
* Nhược điểm:
* Dễ bị ảnh hưởng bởi các giá trị ngoại lệ
* Bị hạn chế khi dữ liệu không tuân theo các giả định về tuyến tính, phân phối phần dư và tính độc lập giữa các biến

### 3.2 Lasso

* Lasso là một thuật toán hồi quy tuyến tính được sử dụng để thực hiện cả việc ước lượng các tham số hồi quy và lựa chọn biến số
* Lasso là kĩ thuật hồi quy áp dụng kỹ thuật hiệu chuẩn để tránh hiện tượng quá khớp bằng cách thêm một phần phạt vào hàm mất mát của hồi quy tuyến tính làm cho các hệ số wi không trở nên quá lớn
* Công thức



Trong đó:

* y là biến phụ thuộc (đầu ra).
* X1,X2,...,Xp​ là các biến độc lập (đặc trưng).
* β0 là hệ số chặn (intercept).
* β1,β2,...,βp là các hệ số hồi quy.
* ϵ là thành phần nhiễu hoặc sai số.
* Ưu điểm:

+ Ngăn ngừa quá khớp( Overfitting)

+ Xử lý tốt với dữ liệu có nhiều chiều

+ Có khả năng loại bỏ các biến không quan trọng

* Nhược điểm

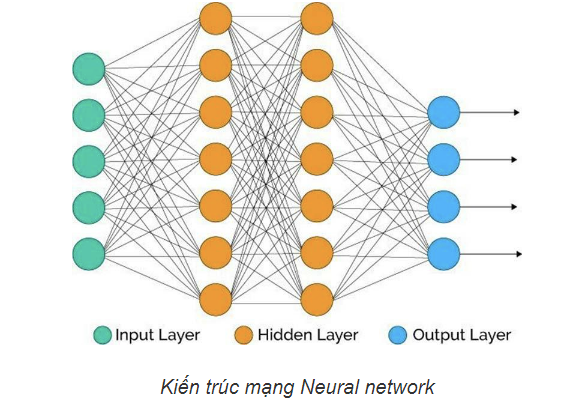
+ Hiệu suất kém với dữ liệu nhỏ hoặc thưa

+ Không xử lý tốt với các biến tương quan cao

+Có thể làm sai lệch hệ số hồi quy

### 3.3 Neural Network

* Neural network là thuật toán được sử dụng trong cả 2 phương pháp học có giám sát và học không giám sát
* Cấu trúc cơ bản:
* Neurons : là đơn vị cơ bản của mạng nơ-ron
* Layers:
* Input layer ( tầng đầu vào): Nhận dữ liệu đầu vào
* Hidden layers ( Tầng ẩn) : Xử lý dữ liệu thông qua các phép toán
* Output Layer ( Tầng đầu ra): Cung cấp kết quả cuối cùng cho nơ ron

******

* Ưu điểm:

+ Khả năng học từ các mẫu dữ liệu phi tuyến và phức tạp

+ Thích hợp cho các bài toán lớn với nhiều đặc trưng hoặc cấu trúc dữ liệu phức tạp

+ Có thể dễ dàng mở rộng và kết hợp các loại mạng khác nhau (CNN, RNN)

* Nhược điểm:

+ Cần dữ liệu lớn để đạt hiệu quả cao

+ Khó giải thích và phân tích, dễ gặp vấn đề với "hộp đen" (black-box)

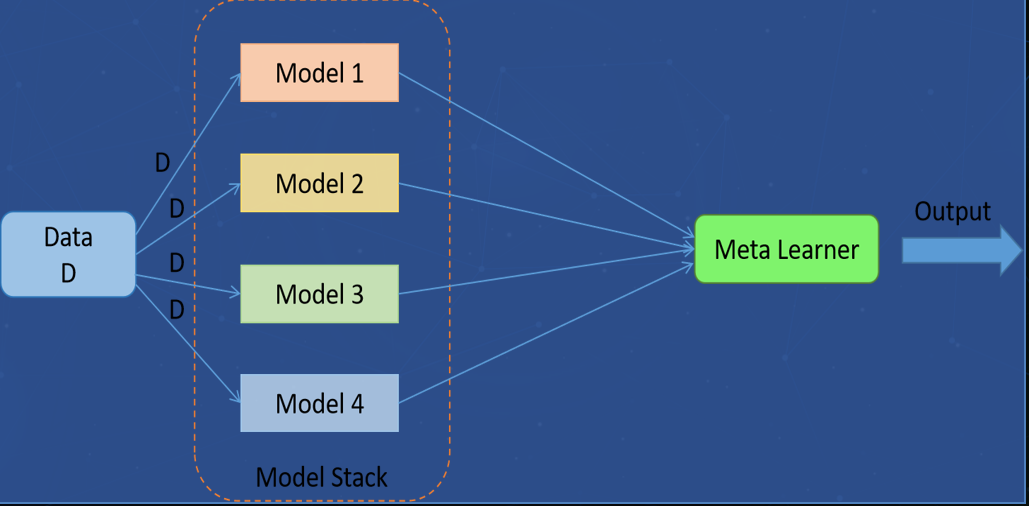
+ Yêu cầu tài nguyên tính toán mạnh và thời gian huấn luyện dài

### 3.4 Stacking

* Stacking là một kỹ thuật ensemble learning , trong đó nhiều mô hình khác nhau được huấn luyện và kết hợp lại để đưa ra dự đoán cuối cùng
* *Quá trình Stacking*

+ Base models ( Mô hình nền tảng): là các mô hình ban đầu có thể là Linear Regression, Lasso và Neural Network

+ Meta-Model (Mô hình bậc hai): Mô hình này sẽ học từ đầu ra của các base models kết hợp các dự đoán từ các mô hình nền tảng và đưa ra dự đoán cuối cùng

****

* Ưu điểm:

+ Tăng hiệu suất

+ Cải thiện độ chính xác và làm giảm overfitting

+ Tận dụng nhiều loại mô hình

* Nhược điểm:

+ Tốn thời gian huấn luyện

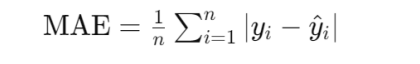
+ Rủi ro overfitting nếu không cẩn thận

## Các phương pháp đánh giá mô hình học máy cho đề tài

### 4.1 MAE ( Mean Absolute Error) : Trung bình lỗi tuyệt đối

*\* Định nghĩa*: MAE là giá trị trung bình của sai số tuyệt đối giữa các giá trị dự đoán

*\* Công thức :*



*\* Tập giá trị :* [0,∞) MAE càng nhỏ thì mô hình càng chính xác

*\* Ý nghĩa:* MAE cho biết sai số trung bình mà mô hình dự đoán sai lệch so với giá trị thực.

*\* Tính chất:*

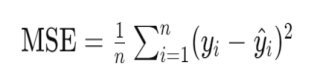
+ MAE dễ hiểu và trự quan vì nó sử dụng sai số tuyệt đối

+ Không bị ảnh hưởng bởi các giá trị ngoại lai lớn , do đó phù hợp với các bài toán mà độ lớn sai số không quan trọng

### 4.2 MSE (Mean Squared Error) - Trung bình bình phương sai số

*\* Định nghĩa:* MSE là giá trị trung bình của bình phương sai số giữa các giá trị thực và giá trị dự đoán.

*\* Công thức:*



*\* Tập giá trị :* [0,∞) MAE càng nhỏ thì mô hình càng chính xác

*\* Ý nghĩa:* MSE cho biết sai số trung bình nhưng phần phạt nặng hơn các sai số lớn do việc bình phương các sai số

*\* Tính chất:*

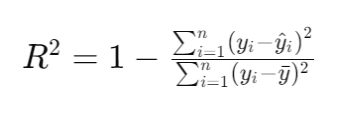
+ MSE nhạy cảm với các giá trị ngoại lai (outliers), vì các sai số lớn sẽ được bình phương, làm tăng giá trị lỗi. Do đó, nếu có nhiều giá trị dự đoán chênh lệch lớn, MSE sẽ bị ảnh hưởng mạnh.

+ Thích hợp khi chúng ta muốn giảm thiểu các sai số lớn và tránh được những sai lệch lớn trong dự đoán.

### 4.3 R2 ( R-squared) : Hệ số xác định

**\*** *Định nghĩa:* R² là tỷ lệ giữa tổng phương sai mà mô hình giải thích được và tổng phương sai của dữ liệu thực tế.

\* Công thức :



*\* Tập giá trị :* R2 có giá trị [0; 1]

+ R2 = 1 : Mô hình dự đoán hoàn hảo

+ R2 = 0: Mô hình dự đoán kém không thể dự đoán tốt hơn giá trị trung bình của tập dữ liệu

*\* Ý nghĩa:* R² đo lường mức độ mà mô hình có thể giải thích được sự biến thiên của dữ liệu. Nó cho biết tỷ lệ phần trăm biến thiên của dữ liệu đầu vào mà mô hình có thể dự đoán chính xác.

*\* Tính chất:*

+ Hữu ích khi so sánh các mô hình khác nhau về khả năng giải thích dữ liệu, nhưng không cho biết mức độ sai số cụ thể như MAE hay MSE.

# CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM:

Trong phần này, chúng em sử dụng các thuật toán Linear Regression, Lasso Regression, MLP và Stacking(kết hợp của 3 mô hình trên để xây dựng mô hình. Sau đó,chúng em đánh giá mức độ tốt của mô hình thông qua 3 hàm đánh giá: MSE, MAE, R2 Score.

## 

## 3.1. Dữ liệu

***3.1.1. Nguồn dữ liệu***

Bộ dữ liệu được lấy từ trang web: <https://www.kaggle.com/code/ybifoundation/simple-linear-regression/input?scriptVersionId=121443641>.

### 3.1.2 Độ tin cậy của dữ liệu

Bộ dữ liệu có tới hơn 1500 người tải về và sử dụng nên có độ tin cậy cao.Vì vậy, bọn em quyết đính sử dụng bộ dữ liệu này để xây dựng mô hình.

### 3.1.3. Các thuộc tính được thu thập

Tập dữ liệu có 1 thuộc tính đó là : Experience Years (Năm kinh nghiệm).

### 3.1.4. Tiền xử lí dữ liệu

Sau khi thu thập dữ liệu bọn em đã kiểm tra những dữ liệu rỗng và những dữ liệu bị trùng lặp. Sau đó, loại bỏ chúng ra khỏi tập dữ liệu để mô hình được huấn luyện tốt hơn và tránh gây nhiễu mô hình.

### 3.1.5. Chia dữ liệu

* Ta chia tập dữ liệu thành phần:

+Thuộc tính: Experience Years.

+Nhãn: Salary.

* Tập dữ liệu sau đó sẽ được chia thành 2 tập con:

+Tâp train: chiếm 80% tâp dữ liệu.

+Tập test: chiếm 20% tập dữ liệu.

## 3.2. Kết quả thực nghiệm với Linear Regression

* Đầu tiên, em sẽ tiến hành xây dựng mô hình dựa trên tập train. Sau đó, dùng tập test và các hàm đánh giá MSE,MAE,R2 để kiểm tra độ tốt của mô hình:

+MAE: 6419.91

+MSE: 48077731.17

+R2: 0.91

## 3.3. Kết quả thực nghiệm với Lasso Regression

* Để mô hình tránh bị overfitting,chúng em sẽ sử dụng mô hình Lasso Regression để dư đoán tham số alpha được thử từ 1-17,,1200,1300,1400 thông qua thuật toán GridSearch kết hợp với cross-validation (với cv = 5) và với điểm đánh giá thông qua mean\_squared\_error (MSE) để tìm ra tham số alpha tốt nhất cho mô hình.
* Kết quả:
* MAE: 6523.18
* MSE: 49502991.91
* R2: 0.90

## 3.4. Kết quả thực nghiệm với Neural Network

* Để xây dựng mô hình, em truyền tham số đầu vào là:

+Hidden\_layer\_sizes = (50,50,50) (mô hình có 3 lớp ẩn và mỗi lớp chứa 50 nơ-ron.

* Sau khi thực thi mô hình, em được kết quả:
* MAE: 5030.95
* MSE: 31726492.88
* R2: 0.94

## 3.5. Kết quả thực nghiệm với Stacking

* Stacking là sự kết hợp của 3 mô hình trên để tạo ra 1 mô hình mới mạnh mẽ hơn. Với các base-model là MLPRegressor, LassoRegression và meta-model là LinearRegression.
* Sau khi chọn được mô hình base-model chúng sẽ được đào tạo thông qua tập dữ liệu train. Sau đó, meta-model sẽ lấy các dự đoán của base-model làm đầu vào và đưa ra dự đoán cuối cùng.
* Kết quả:
* MAE: 5027.82
* MSE: 31846977.32
* R2: 0.94

Nhận xét: Thông qua các mô hình trên, em có thể thấy mô hình Stacking đưa ra các dự đoán tốt nhất với R2 lớn nhất và MAE,MSE nhỏ nhất.

+30460880.2830460880.2830460880.2830460880.28

# PHẦN 4: KẾT LUẬN VÀ ĐÁNH GIÁ

## Đánh giá

* Kết quả đánh giá đã rõ ràng chỉ ra mô hình Stacking đem lại độ chính xác tốt nhất trong 4 mô hình Linear Regression, Lasso Regression, Neural Network và Stacking.
* Vì vậy, với dữ liệu từ 40 mẫu để có một dự báo chính xác tính cách thì ta nên lựa chọn sử dụng dự báo tính cách bằng mô hình Stacking Regression.

## Kết luận

* Kết quả nghiên cứu với đề tài "**Dự đoán lương dựa trên số năm kinh nghiệm**" đã mang lại một số kết quả đáng chú ý cũng như phản ánh một số hạn chế:

### 2.1 Kết quả đạt được:

- Hiểu biết tổng quan về vấn đề dự báo lương.

- Nghiên cứu và áp dụng kỹ thuật học máy (Machine Learning) vào việc dự báo lương dựa trên số năm kinh nghiệm.

- Thử nghiệm và đánh giá hiệu suất của mô hình sử dụng các thuật toán như Linear Regression, Lasso Regression, Neural Network và Stacking.

### 2.2 Những hạn chế:

- Hạn chế chỉ có 1 đặc trưng đầu vào là số năm kinh nghiệm, không tránh khỏi một số sai sót trong quá trình dự đoán.

- Dữ liệu cho việc huấn luyện mô hình có thể còn hạn chế, không cho ra kết quả dự đoán tốt nhất trong một vài trường hợp.

* Mô hình dự đoán tính cách này có thể được sử dụng trong phân phối lương cho nhân viên, hay nhân viên có thể dự đoán lương của chính mình dựa vào số năm kinh nghiệm làm việc của mình. Vì vậy, dự đoán lương được hệ thống dự đoán sau khi người dùng thực hiện khảo sát bằng tập dữ liệu được cung cấp. Phân tích lương và dự đoán có thể được dùng nhiều hơn trong tương lai có thể bổ sung thêm nhiều đầu vào để dự đoán lương chính xác hơn . Hơn nữa, mọi cải tiến đều có thể được thực hiện bằng cách sử dụng tập dữ liệu và thuật toán để cải thiện độ chính xác và đánh giá trên hàm đánh giá để có thể có thể có cái nhìn tổng quan hơn và có thể hữu ích cho việc dự đoán lương.

# TÀI LIỆU THAM KHẢO

[1] Simple Linear Regression

<https://www.kaggle.com/code/ybifoundation/simple-linear-regression/notebook>

[2] scikit-learn. Lasso

<https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.Lasso.html>

[3] scikit-learn. MLPRegressor<https://scikit-learn.org/dev/modules/generated/sklearn.neural_network.MLPRegressor.html>

[4] Vũ Hữu Tiệp. Linear Regression. Machine Learning cơ bản.

<https://machinelearningcoban.com/2016/12/28/linearregression/>

[5] Loss function

<https://www.geeksforgeeks.org/ml-common-loss-functions/>