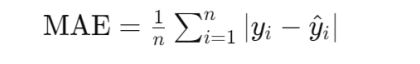
**Technical report (nhóm 4 lớp 64ktpm4)**

1. **Các hàm đánh giá trong đề tài**
2. **MAE ( Mean Absolute Error) : Trung bình lỗi tuyệt đối**

*\* Định nghĩa*: MAE là giá trị tring 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:*

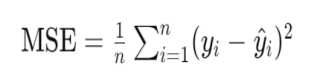
+ MAE dễ hiểu và trự quan vì nó sử dụng sai số tuyệt đối

+ 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

### 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:*

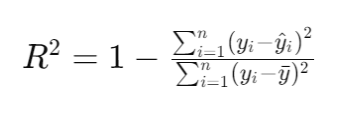
+ MSE nhạy cảm với các giá trị ngoại lai (outliers), vì các sai số lớn sẽ được bình phương, làm tăng giá trị lỗi. Do đó, nếu có nhiều giá trị dự đoán chênh lệch lớn, MSE sẽ bị ảnh hưởng mạnh.

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

**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.

1. **So sánh 3 tham số đánh giá**

\* **MAE** cung cấp cái nhìn trực quan về sai số trung bình, nhưng không phạt mạnh các giá trị ngoại lai, nên thích hợp trong các trường hợp không muốn mô hình bị ảnh hưởng bởi các lỗi lớn.

\* **MSE** phạt nặng các sai số lớn, giúp chúng ta tối ưu hóa mô hình theo hướng giảm thiểu sai số lớn, nhưng có thể làm tăng độ nhạy cảm với các giá trị ngoại lai.

\* **R²** không trực tiếp cho biết sai số dự đoán, mà đo lường mức độ mô hình giải thích được sự biến thiên trong dữ liệu. Nó hữu ích khi so sánh độ phù hợp của các mô hình khác nhau, nhưng không cho biết rõ ràng mức độ dự đoán chính xác.

1. **Các tham số truyền vào các phương pháp**
2. **Giới thiệu:**

* Báo cáo này giải thích các tham số và tham số quan trọng được truyền vào hai mô hình học máy: **Neural network** (mô hình dựa trên mạng nơ-ron) và **Lasso** (mô hình tuyến tính với regularization L1). Hiểu rõ các tham số này là điều quan trọng để tối ưu hóa mô hình, đảm bảo hiệu suất và khả năng phù hợp với dữ liệu.

1. **Tham số ở phương pháp neural network:**

* hidden\_layer\_sizes=(50,50,50)

**Ý nghĩa**: Xác định số lượng lớp ẩn và số lượng nơ-ron trong mỗi lớp ẩn của mạng nơ-ron. Ở đây, có 3 lớp ẩn, mỗi lớp có 50 nơ-ron. Cấu hình này giúp mô hình có thể nắm bắt các mối quan hệ phức tạp giữa các đặc trưng đầu vào, làm cho nó phù hợp với dữ liệu có dạng phi tuyến tính.

**Ảnh hưởng đến đầu ra**:

* Số lượng nơ-ron và lớp ẩn ảnh hưởng trực tiếp đến khả năng học của mô hình. Cấu trúc phức tạp hơn (nhiều lớp ẩn và nơ-ron) sẽ giúp mô hình học được các mối quan hệ phi tuyến tính trong dữ liệu.
* Tuy nhiên, nếu số lớp và số nơ-ron quá lớn, mô hình có thể bị overfitting, dẫn đến việc dự đoán tốt trên tập huấn luyện nhưng kém trên tập kiểm thử.
* Với 3 lớp ẩn, mô hình có khả năng học tốt các mối quan hệ phức tạp.

1. **Tham số truyền vào phương pháp Lasso Regression:**

* alpha=1200

 **Ý nghĩa**: Đây là tham số regularization, điều chỉnh mức độ phạt lên các trọng số lớn. Alpha cao sẽ áp đặt hình phạt lớn lên các trọng số lớn, giảm overfitting bằng cách giảm giá trị trọng số, thậm chí đưa một số trọng số về 0. Thông qua thư viện GridSearchCV ta tìm được alpha tốt nhất là 1200

 **Ảnh hưởng đến đầu ra**:

* Với alpha=1200, mô hình Lasso sẽ mạnh tay hơn trong việc đẩy các trọng số không quan trọng về 0. Điều này giúp mô hình đơn giản hơn và tránh overfitting.
* Tuy nhiên, nếu alpha quá lớn, mô hình có thể dẫn đến underfitting, vì quá nhiều trọng số bị giảm hoặc triệt tiêu.

1. **Kết luận:**

* Neural network:Là một mô hình mạng nơ-ron với ba lớp ẩn, mỗi lớp có 50 neuron, phù hợp cho các tác vụ hồi quy phức tạp với dữ liệu phi tuyến tính.
* **Lasso Regression**: Là một mô hình tuyến tính với regularization L1 (alpha=1200), với ít hệ số khác 0, đơn giản hóa và làm mô hình dễ diễn giải hơn.
* Trong cả hai mô hình, các siêu tham số được chọn đóng vai trò quan trọng trong việc quyết định hành vi và hiệu suất của mô hình. Với neural network, kiến trúc của các lớp ẩn quyết định độ phức tạp của mô hình, trong khi ở lasso, độ mạnh của regularization (alpha) quyết định sự cân bằng giữa độ chính xác và độ đơn giản của mô hình.

1. **SỰ THAY ĐỔI KẾT QUẢ KHI THÊM BỚT DỮ LIỆU:**

**Kết quả ban đầu (khi chưa bớt dữ liệu)**

**- Neuralnetwork:**

+ MSE: 17574906.25

+ MAE: 3539.61

+ R2: 0.97

**-Lasso Regression:**

+ MSE: 34257606.84

+ MAE: 4886.01

+R2: 0.95

**-Linear Regression:**

+MSE: 35997586.39

+MAE: 5027.74

+R2: 0.95

**-Stacking:**

+MSE: 19113700

+MAE: 5027.74

R2: 0.95

**Kết quả sau khi số lượng dữ liệu đi 25% (từ 40 dữ liệu xuống 30 dữ liệu):**

**- Neuralnetwork:**

+ MSE: 23811876.86

+ MAE:4112.14

+ R2:0.92

**-Lasso Regression:**

+ MSE: 51264789.28

+ MAE: 7015.83

+R2: 0.82

**-Linear Regression:**

+MSE: 50840047.11

+MAE: 0.82

+R2: 0.98

**-Stacking:**

+MSE: 22621493.11

+MAE: 4301.30

R2: 0.92

-Quan sát thấy MSE và MAE tăng sau khi bớt dữ liệu và R2 giảm .

**Kết luận:**

### 1. ****Thêm dữ liệu****

* **Cải thiện độ chính xác**: Khi thêm nhiều dữ liệu, đặc biệt là dữ liệu đa dạng và đầy đủ, mô hình có thể học được nhiều hơn về các mối quan hệ phức tạp trong tập dữ liệu, dẫn đến cải thiện độ chính xác.
* **Giảm overfitting**: Overfitting xảy ra khi mô hình quá phù hợp với tập huấn luyện và không thể tổng quát hóa tốt trên tập dữ liệu chưa thấy. Thêm dữ liệu thường giúp mô hình tổng quát hóa tốt hơn, giảm tình trạng overfitting.
* **Tăng thời gian huấn luyện**: Khi thêm dữ liệu, đặc biệt là dữ liệu lớn, thời gian huấn luyện mô hình sẽ tăng lên, đòi hỏi nhiều tài nguyên tính toán hơn.
* **Cải thiện tính ổn định**: Nhiều dữ liệu hơn giúp mô hình có thể tạo ra các dự đoán ổn định hơn, ít bị ảnh hưởng bởi các mẫu nhiễu hoặc bất thường trong tập dữ liệu.

### 2. ****Bớt dữ liệu****

* **Giảm độ chính xác**: Nếu dữ liệu bị bớt đi, mô hình có thể giảm độ chính xác.
* **Tăng khả năng overfitting**: Với ít dữ liệu hơn, mô hình có xu hướng ghi nhớ quá mức các đặc trưng của tập huấn luyện và không thể tổng quát hóa tốt khi gặp dữ liệu mới.
* **Giảm thời gian huấn luyện**: Việc bớt dữ liệu sẽ giảm thời gian huấn luyện, tiết kiệm tài nguyên tính toán.
* **Giảm hiệu suất trên dữ liệu không thấy**: Nếu tập dữ liệu quá nhỏ hoặc không đủ đa dạng, mô hình có thể không hiệu quả khi được áp dụng trên các tập dữ liệu mới, lạ.

### 3. ****Chất lượng của dữ liệu****

Thêm hoặc bớt dữ liệu không chỉ phụ thuộc vào số lượng mà còn phụ thuộc vào chất lượng của dữ liệu. Nếu dữ liệu thêm vào là dữ liệu nhiễu hoặc không phù hợp, mô hình có thể học được các đặc trưng sai và làm giảm hiệu suất.

Tóm lại, việc thêm dữ liệu thường mang lại lợi ích nhưng phải cân nhắc về chất lượng và sự đa dạng của dữ liệu. Bớt dữ liệu có thể giảm thời gian huấn luyện nhưng cũng đi kèm với nguy cơ làm giảm độ chính xác và khả năng tổng quát hóa của mô hình.