**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**LỚP: 21CLC07**

**Ảnh có chứa Phông chữ, biểu tượng, Đồ họa, Xanh điện

Mô tả được tạo tự động**

**BÁO CÁO PROJECT 3:**

**LINER REGRESSION**

Thông tin sinh viên thực hiện:

* Họ và Tên: Nguyễn Tấn Khiêm
* MSSV: 21127726

Giảng viên hướng dẫn:

• Giảng viên: Nguyễn Văn Quang Huy  
• Giảng viên: Ngô Đình Hy  
• Giảng viên: Trần Hà Sơn  
• Giảng viên: Nguyễn Đình Thúc

1. **Đánh giá mức độ hoàn thành:**

|  |  |  |
| --- | --- | --- |
| STT | Nội dung | Mức độ hoàn thành |
| 1 | 1a. Sử dụng 11 đặc trưng đầu tiên đề bài cung cấp. | 100% |
| 2 | 1b. Phân tích ảnh hưởng của ‘đặc trưng tính cách’ dựa trên điểm các bài kiểm tra của AMCAT. | 100% |
| 3 | 1c. Phân tích ảnh hưởng của ‘đặc trưng ngoại ngữ, lô-gic, định lượng’ đến mức lương của các kỹ sư dựa trên điểm các bài kiểm tra của AMCAT | 100% |
| 4 | 1d. Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất. | 100% |
| 5 | Báo cáo về kết quả, đánh giá và nhận xét các mô hình đã xây dựng | 100% |

1. **Khai báo thư viện và các hàm sử dụng từ thư viện đó:**

**1. pandas:**

- Mô tả: Thư viện Pandas cung cấp các cấu trúc dữ liệu và công cụ phân tích dữ liệu dễ sử dụng và hiệu quả trong Python.

- Hàm gọi:

- `pd.read\_csv(file\_path)`: Đọc dữ liệu từ tệp CSV và trả về DataFrame.

- `pd.concat(objs, ignore\_index)`: Nối các đối tượng (DataFrame, Series) theo chiều dọc.

**2. numpy:**

- Mô tả: Thư viện NumPy cung cấp hỗ trợ cho việc làm việc với mảng và ma trận trong Python, bao gồm các phép toán và chức năng số học.

- Hàm gọi:

- `np.array(iterable)`: Chuyển một iterable (danh sách, tuple, ...) thành mảng NumPy.

- `np.mean(arr)`: Tính giá trị trung bình của mảng.

**3. math:**

- Mô tả: Thư viện math cung cấp các chức năng toán học cơ bản.

- Hàm gọi:

- `math.sqrt(number)`: Tính căn bậc hai của một số.

**4. time:**

- Mô tả: Thư viện time cung cấp các chức năng liên quan đến thời gian.

- Hàm gọi:

- `time.time()`: Trả về thời gian hiện tại dưới dạng số thời gian Unix.

**5. sklearn.linear\_model.LinearRegression:**

- Mô tả: Thư viện scikit-learn (sklearn) cung cấp các công cụ cho việc xây dựng và đào tạo các mô hình học máy.

- Hàm gọi:

- `LinearRegression()`: Tạo một mô hình hồi quy tuyến tính.

- `fit(X, y)`: Đào tạo mô hình trên tập dữ liệu X và nhãn y.

- `predict(X)`: Dự đoán đầu ra cho các mẫu trong tập dữ liệu X.

**6. sklearn.model\_selection.cross\_val\_score:**

- Mô tả: Thư viện sklearn cung cấp các công cụ cho việc thực hiện cross-validation.

- Hàm gọi:

- `cross\_val\_score(estimator, X, y, cv, scoring)`: Thực hiện k-fold cross-validation trên mô hình với tập dữ liệu X, nhãn y và số lượng phân chia cv.

**7. sklearn.metrics.mean\_absolute\_error:**

- Mô tả: Thư viện sklearn cung cấp các công cụ tính toán các chỉ số đánh giá mô hình.

- Hàm gọi:

- `mean\_absolute\_error(y\_true, y\_pred)`: Tính Mean Absolute Error giữa giá trị dự đoán y\_pred và giá trị thực tế y\_true.

**8. itertools.combinations:**

- Mô tả: Thư viện itertools cung cấp các công cụ để tạo ra các tổ hợp từ một tập hợp.

- Hàm gọi:

- `combinations(iterable, r)`: Trả về tất cả các tổ hợp có r phần tử từ iterable.

1. **Mô tả hàm tự tạo:**

**1. alculate\_mae\_for\_feature\_combination(data, feature\_combination):**

- Mô tả: Hàm này tính giá trị trung bình của Mean Absolute Error (MAE) bằng cách sử dụng k-fold cross-validation cho một tổ hợp cụ thể của các đặc trưng trên tập dữ liệu.

- Đối số:

- data: DataFrame chứa các đặc trưng và cột 'Salary'.

- feature\_combination: Tổ hợp các đặc trưng cần kiểm tra.

- Trả về: Giá trị trung bình của MAE.

**2. create\_linear\_formula(coefficients, intercept, feature\_names):**

- Mô tả: Hàm này tạo chuỗi biểu diễn công thức hồi quy tuyến tính dựa trên các hệ số, hệ số chặn và tên các đặc trưng.

- Đối số:

- coefficients: Mảng các hệ số hồi quy.

- intercept: Hệ số chặn.

- feature\_names: Danh sách tên các đặc trưng.

- Trả về: Chuỗi biểu diễn công thức hồi quy tuyến tính.

**3. find\_best\_feature\_combination(data, features, num\_features):**

- Mô tả: Hàm này tìm tổ hợp tốt nhất của các đặc trưng bằng cách thử tất cả các tổ hợp có thể và tính MAE cho mỗi tổ hợp.

- Đối số:

- data: DataFrame chứa các đặc trưng và cột 'Salary'.

- features: Danh sách tất cả các đặc trưng có thể.

- num\_features: Số lượng đặc trưng cần chọn.

- Trả về: Tổ hợp tốt nhất của các đặc trưng, MAE tốt nhất và công thức hồi quy tuyến tính tương ứng.

**4. model(num\_features):**

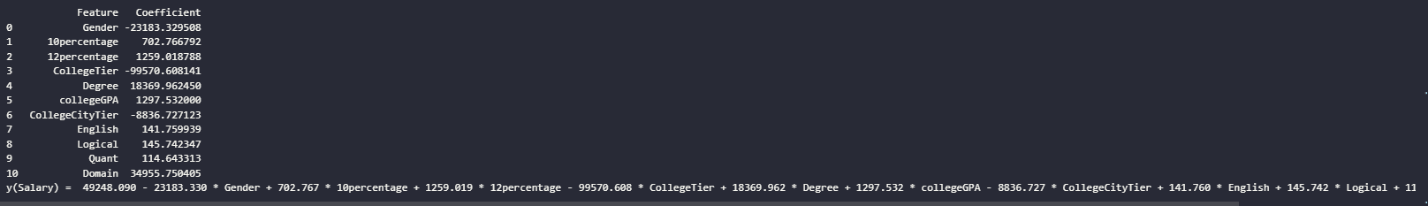
- Mô tả: Hàm này chạy quá trình tìm tổ hợp tốt nhất của đặc trưng cho một số lượng đặc trưng đã cho.

- Đối số:

- num\_features: Số lượng đặc trưng cần chọn.

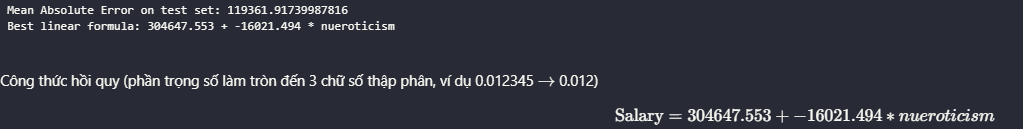
- Trả về: Tổ hợp tốt nhất của các đặc trưng, MAE tốt nhất và công thức hồi quy tuyến tính tương ứng.

1. **Báo cáo kết quả từ mô hình xây dựng:**
2. **Sử dụng 11 đặc trưng đầu tiên đề bài cung cấp:**



1. **Phân tích ảnh hưởng của ‘đặc trưng tính cách’ dựa trên điểm các bài kiểm tra của AMCAT:**

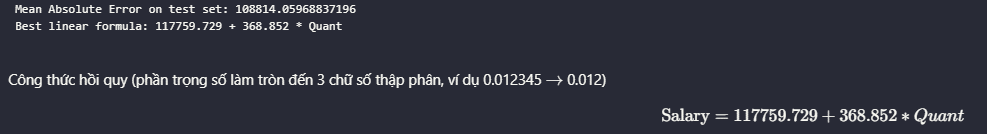
Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động 

1. **Phân tích ảnh hưởng của ‘đặc trưng ngoại ngữ, lô-gic, định lượng’ đến mức lương của các kỹ sư dựa trên điểm các bài kiểm tra của AMCAT:**

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, số

Mô tả được tạo tự động



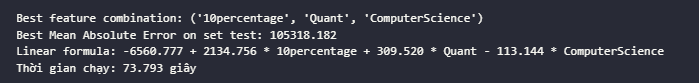
1. **Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất.**

* Mô hình 1:

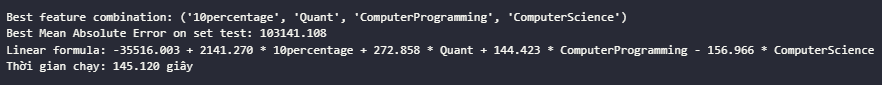
Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình

Mô tả được tạo tự động

* Mô hình 2:



* Mô hình 3:



* Báo cáo kết quả:

|  |  |  |
| --- | --- | --- |
| STT | Mô hình | MAE |
| 1 | Sử dụng 2 đặc trưng | 106022.179 |
| 2 | Sử dụng 3 đặc trưng | 105318.182 |
| 3 | Sử dụng 4 đặc trưng | 103141.182 |

* Cách tìm ra mô hình tốt nhất:

+ Đầu tiên, với môi hình ta sẽ tìm đặc trưng tốt nhất cho mô hình đó bằng cách tạo các tổ hợp từ toàn bộ đặc trưng của đề bài, sau đó chạy toàn bộ các tổ hợp đó để tìm đặc trưng tốt nhất đối với mô hình.

+ Sau khi tìm được các đặc trưng tốt nhất cho mỗi mô hình thì ta so sánh các mô hình với nhau xem mô hình nào có MAE thấp nhất.

+ Tìm Linear formula từ mô hình tốt nhất và các đặc trưng tốt nhất cho mô hình đó.

* Kết quả từ mô hình tốt nhất:

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động