BÁO CÁO ĐỒ ÁN

Giảng viên hướng dẫn: Thầy Trần Trung Kiên

Nhóm 1:

18120529 – Phan Văn Võ Quyền

18120540 – Phạm Minh Sỹ

I. THU THẬP DỮ LIỆU

- Hình thức thu thập dữ liệu: API
- Link dữ liệu: FoodData Central (usda.gov)
- Mô tả: dữ liệu trình bày chi tiết về thành phần dinh dưỡng của món ăn thông qua tên thành phần cùng định lượng của nó
- Lí do lựa chọn dữ liệu: nguồn cung cấp thực phẩm và sự hiểu biết khoa học về mối quan hệ giữa khẩu phần ăn và sức khỏe đã phát triển qua nhiều năm do nhu cầu về sức khỏe của con người ngày càng cao

HÌNH ẢNH VỀ DỮ LIỆU THU THẬP ĐƯỢC

Out[4]:

	fdcId	description	dataType	publicationDate	foodCode	foodNutrients	ndbNumber
0	1104067	100 GRAND Bar	Survey (FNDDS)	2020-10-30	91715300.0	[{'number': '203', 'name': 'Protein', 'amount': 2.5, 'unitName': 'G'}, {'number': '204', 'name': 'Total lipid (fat)', 'amount': 19.3, 'unitName': 'G'}, {'number': '205', 'name': 'Carbohydrate, by difference', 'amount': 71.0, 'unitName': 'G'}, {'n...	NaN
1	1104086	3 MUSKETEERS Bar	Survey (FNDDS)	2020-10-30	91726420.0	[{'number': '203', 'name': 'Protein', 'amount': 2.6, 'unitName': 'G'}, {'number': '204', 'name': 'Total lipid (fat)', 'amount': 12.8, 'unitName': 'G'}, {'number': '205', 'name': 'Carbohydrate, by difference', 'amount': 77.8, 'unitName': 'G'}, {'n...	NaN
2	1104087	3 Musketeers Truffle Crisp Bar	Survey (FNDDS)	2020-10-30	91726425.0	[{'number': '203', 'name': 'Protein', 'amount': 6.41, 'unitName': 'G'}, {'number': '204', 'name': 'Total lipid (fat)', 'amount': 28.8, 'unitName': 'G'}, {'number': '205', 'name': 'Carbohydrate, by difference', 'amount': 63.2, 'unitName': 'G'}, {'...	NaN
3	1099098	Abalone, cooked, NS as to cooking method	Survey (FNDDS)	2020-10-30	26301110.0	[{'number': '203', 'name': 'Protein', 'amount': 20.4, 'unitName': 'G'}, {'number': '204', 'name': 'Total lipid (fat)', 'amount': 4.59, 'unitName': 'G'}, {'number': '205', 'name': 'Carbohydrate, by difference', 'amount': 7.26, 'unitName': 'G'}, {'...	NaN
4	1099099	Abalone, floured or breaded, fried	Survey (FNDDS)	2020-10-30	26301140.0	[{'number': '203', 'name': 'Protein', 'amount': 18.2, 'unitName': 'G'}, {'number': '204', 'name': 'Total lipid (fat)', 'amount': 11.4, 'unitName': 'G'}, {'number': '205', 'name': 'Carbohydrate, by difference', 'amount': 15.5, 'unitName': 'G'}, {'...	NaN
5	1099100	Abalone, steamed or poached	Survey (FNDDS)	2020-10-30	26301160.0	[{'number': '203', 'name': 'Protein', 'amount': 34.0, 'unitName': 'G'}, {'number': '204', 'name': 'Total lipid (fat)', 'amount': 1.51, 'unitName': 'G'}, {'number': '205', 'name': 'Carbohydrate, by difference', 'amount': 12.0, 'unitName': 'G'}, {'...	NaN
6	167782	Abiyuch, raw	SR Legacy	2019-04-01	NaN	[{'number': '318', 'name': 'Vitamin A, IU', 'amount': 100, 'unitName': 'IU', 'derivationCode': 'A', 'derivationDescription': 'Analytical'}, {'number': '268', 'name': 'Energy', 'amount': 290, 'unitName': 'kJ', 'derivationCode': 'NC', 'derivationDe...	9427.0

II. KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

- Số dòng, cột của dữ liệu: (10 000, 7)
- Số dòng dữ liệu bị thiếu: 0
- Các cột dữ liệu bị thiếu:
  - fdcId: 0
  - description: 0
  - dataType: 0
  - publicationDate: 0
  - foodCode: 5402
  - foodNutrients: 0
  - ndbNumber: 4603

II. KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

- **Ý nghĩa của mỗi dòng dữ liệu:**
  - Kết quả phân tích thành phần dinh dưỡng của món ăn
- **Ý nghĩa của các cột dữ liệu:**
  - fdcld: chuỗi số duy nhất dùng để định danh các loại món ăn được quy định bởi FCD
  - nbdNumber: chuỗi số duy nhất được dùng để định danh các loại món ăn được quy định bởi SR Legacy foods khác với fcdld
  - description: mô tả loại món ăn, bao gồm tên và đặc trưng món ăn
  - datatype: thông tin loại khảo sát tiến hành
  - publicationDate: thời gian công bố dữ liệu phân tích về món ăn
  - foodCode: chuỗi số duy nhất được dùng để định danh các loại món ăn được quy định bởi FNDDS
  - FoodNutrients: kết quả phân tích chi tiết về thành phần dinh dưỡng

TẠO CỘT THUỘC TÍNH MỚI TỪ CỘT THUỘC TÍNH CŨ

- Chú ý đến cột thuộc tính quan trọng ban nãy:
  - FoodNutrients: kết quả phân tích chi tiết về thành phần dinh dưỡng
- Cần thực hiện:
  - Thực hiện phân tách cột thuộc tính này thành các thuộc tính thành phần dựa theo tên thành phần và đơn vị đo lường
  - Đồng thời tính toán định lượng của thành phần đó
  - Sau phân tách thu được dataframe với fdcId là index, các cột còn lại biểu diễn thành phần dinh dưỡng:

fdcId	trans-beta-Carotene (ug)	trans-Lycopene (ug)	cis-beta-Carotene (ug)	cis-Lycopene (ug)	cis-Lutein/Zeaxanthin (ug)	Zinc, Zn (mg)	Zeaxanthin (ug)	Water (g)	Vitamin K (phylloquinone) (ug)	Vitamin K (Menaquinone-4) (ug)	...	14:1 t (g)	14:1 c (g)	14:1 (g)	14:0 (g)	1
1104067	NaN	NaN	NaN	NaN	NaN	0.99	NaN	6.10	6.2	NaN	...	NaN	NaN	NaN	1.030	N
1104086	NaN	NaN	NaN	NaN	NaN	0.55	NaN	5.80	2.9	NaN	...	NaN	NaN	NaN	0.295	N
1104087	NaN	NaN	NaN	NaN	NaN	1.83	NaN	0.18	5.8	NaN	...	NaN	NaN	NaN	1.650	N
1099098	NaN	NaN	NaN	NaN	NaN	0.98	NaN	65.10	31.7	NaN	...	NaN	NaN	NaN	0.028	N
1099099	NaN	NaN	NaN	NaN	NaN	1.06	NaN	52.40	32.0	NaN	...	NaN	NaN	NaN	0.024	N

5 rows × 220 columns

ĐẶT CÂU HỎI?

- Đâu là công thức tính năng lượng từ các thành phần dinh dưỡng cụ thể?
  - Input được xác định là các cột thành phần dinh dưỡng
  - Output được xác định là cột năng lượng
  - Dễ nhận thấy đây là bài toán hồi quy
  - Như vậy năng lượng sẽ được tính thông qua một công thức với các hạng tử là các thành phần dinh dưỡng cụ thể
- Việc tìm ra câu trả lời cho câu hỏi này có ý nghĩa:
  - Từ việc biết được mức năng lượng mà mỗi loại thực phẩm mang lại, con người ta có thể chủ động trong việc điều phối khẩu phần ăn hằng ngày sao cho cân bằng giữa mức năng lượng cần nạp vào và sức khỏe của họ, đặc biệt trong bối cảnh con người ngày càng chú trọng hơn vào sức khỏe của mình
  - Ngoài ra nó còn giúp các chuyên gia, nhà nghiên cứu phát triển các sản phẩm phù hợp thị hiếu sức khỏe mà vẫn đảm bảo năng lượng cùng chi phí sản xuất
- Nguồn cảm hứng của câu hỏi:
  - Phải nói rằng nguồn cảm hứng của câu hỏi xuất phát từ nguồn cảm hứng chọn chủ đề của nhóm em, nhóm em nhận thấy rằng tỉ lệ béo phì ở dân số các quốc gia mỗi ngày một tăng, họ luôn đặt ra câu hỏi tại sao tôi ăn ít nhưng vẫn mập. Tuy nhiên họ không biết rằng họ ăn ít về khối lượng nhưng lượng calo trong thực phẩm họ ăn không hề thấp, do đó nhóm em đã đặt ra câu hỏi là đâu là công thức tính năng lượng từ các thành phần dinh dưỡng của món ăn, nhằm giúp mọi người hiểu được thực phẩm họ ăn chứa nhiều năng lượng tới đâu để có thể điều chỉnh chế độ ăn cân bằng và hợp lý

KHÁM PHÁ VÀ TIỀN XỬ LÝ TRƯỚC KHI TÁCH TẬP

- DataType của cột “Energy (Cal)”: float64
- Số dòng thiếu của cột “Energy (Cal)”: 28
- Thực hiện xóa các dòng dữ liệu thiếu

TÁCH CÁC TẬP DỮ LIỆU

- Tách thành 3 tập train, validation, test với tỉ lệ lần lượt là: 60%, 20%, 20%
- Vector input X: tất cả các cột dữ liệu ngoại trừ cột “Energy (Cal)”
- Output y: cột dữ liệu “Energy (Cal)”

KHÁM PHÁ VÀ TIỀN XỬ LÝ TẬP HUẤN LUYỆN (KHÁM PHÁ)

- Kiểu dữ liệu của các cột trong vector input X: float64
- Tính toán các thông số trên các cột dữ liệu: missing_ratio, lower_quantile, median, upper_quantile

	trans-beta-Carotene (ug)	trans-Lycopene (ug)	cis-beta-Carotene (ug)	cis-Lycopene (ug)	cis-Lutein/Zeaxanthin (ug)	Zinc, Zn (mg)	Zeaxanthin (ug)	Water (g)	Vitamin K (phylloquinone) (ug)	Vitamin K (Menaquinone-4) (ug)	...	14:1 t (g)	14:1 c (g)	14:1 (g)
missing_ratio	100.0	100.0	100.0	100.0	100.0	2.8	100.0	0.0	19.6	95.9	...	99.900	99.60	81.50
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000	0.00	0.00
lower_quartile	0.2	0.0	0.2	0.0	14.2	0.4	57.2	43.0	0.5	0.0	...	0.000	0.00	0.00
median	0.5	0.0	0.5	0.0	28.5	0.8	114.5	64.6	2.2	0.0	...	0.000	0.00	0.00
upper_quartile	0.8	0.0	0.8	0.0	42.8	2.2	171.8	80.1	7.5	1.7	...	0.000	0.00	0.00
max	1.0	0.0	1.0	0.0	57.0	98.9	229.0	100.0	1640.0	35.7	...	0.047	0.31	1.43

6 rows × 219 columns

KHÁM PHÁ VÀ TIỀN XỬ LÝ TẬP HUẤN LUYỆN (TIỀN XỬ LÝ)

- Với các cột có liên quan đến Energy (kcal), Energy chỉ được đo lường bằng kcal từ tháng 10 năm 2020 và để cho đơn giản, ta chỉ xét đến Energy (kcal) nên loại bỏ các cột:
  - Energy (kj)
  - Energy (Atwater Specific Factors) (kcal)
  - Energy (Atwater General Factors) (kcal)
- Với các cột liên quan đến lipit, ta loại bỏ các cột acid béo, giữ lại cột tổng cộng lượng acid béo cuối cùng, giảm bớt nguy cơ overfitting
- Loại bỏ các cột dữ liệu có tỉ lệ thiếu trên 25%, còn lại sử dụng phương pháp mean để điền dữ liệu bị thiếu

KHÁM PHÁ VÀ TIỀN XỬ LÝ TẬP HUẤN LUYỆN (TIỀN XỬ LÝ)

- Xây dựng class ColDropper kế thừa class BaseEstimator và TransformerMixin để thực hiện nhiệm vụ tiền xử lý xóa các cột đã nêu trên
- Xây dựng process_pipeline gồm các giai đoạn:
  - ColDropper(missing_ratio_threshold=0.25): xóa cột với ngưỡng 25% giá trị thiếu
  - SimpleImputer(stragety='mean'): thực hiện điền dữ liệu bị thiếu bằng chiến lược mean
  - StandardScaler(): chuẩn hóa tỉ lệ dữ liệu

  ⇒Sau process_pipeline dữ liệu thu được có kích thước: (5938, 45) -> còn khá nhiều chiều
- Xây dựng process_pipeline_full:
  - Sử dụng PCA() thực hiện cắt giảm chiều dữ liệu sao cho vẫn đảm bảo mối tương quan
  - Thêm vào với process_pipeline tạo thành process_pipeline_full

  ⇒Sau process_pipeline_full dữ liệu thu được có kích thước: (5983, 20)

TIỀN XỬ LÝ VÀ MÔ HÌNH HÓA (VALIDATION)

- Thực hiện transform trên tập validation
- Tiến hành lựa chọn mô hình tốt nhất:
  - Neural Network
  - SVR (Support Vector Regressor)
  - RandomForestRegressor

MÔ HÌNH NEURAL NETWORK

- Thử nghiệm mô hình neural network với:
  - hidden_layer_sizes=(20), activation='tanh', solver='lbfgs', random_state=0, max_iter=2500
  - Siêu tham số alpha của MLPRegressor với 5 giá trị khác nhau: 0.1, 1, 10, 100, 1000
  - Sau khi xây dựng mô hình, tạo một full_pipeline chứa process_pipeline_full và mô hình
  - Tìm được best_alpha=100 và best_val_err=1.103

MÔ HÌNH SVR

- Thử nghiệm mô hình neural network với:
  - max_iter=10 000
  - Siêu tham số C của SVR với 5 giá trị khác nhau: 1, 3, 5, 7, 9 và siêu tham số epsilon với 5 giá trị: 0.1, 0.15, 0.2, 0.25, 0.3
  - Tìm được best_c=9, best_epsilon=0.1 và best_val_err=8.841

MÔ HÌNH RANDOMFORESTREGRESSOR

- Thử nghiệm mô hình RandomForestRegressor với:
  - random_state=0
  - Siêu tham số n_estimator của RandomForestRegressor với 5 giá trị khác nhau: 50, 75, 100, 125, 150 và siêu tham số max_depth với 5 giá trị: 16, 32, 64, 128, 256
  - Tìm được best_num=50, best_depth=256 và best_val_err=2.238

ĐÁNH GIÁ 3 MÔ HÌNH

- Mô hình Neural Network, RandomForest cho ra kết quả khả quan, tỉ lệ đúng trên tập validation cao, trong khi đó mô hình SVR lại không cho kết quả tốt bằng 2 mô hình trên
- Việc lựa chọn giữa 2 mô hình Neural Network và RandomForest lúc này mang ý nghĩa hơi chủ quan vì nó còn phụ thuộc vào siêu tham số ta chọn, đồng thời kết quả hai mô hình đều tốt
- Lựa chọn mô hình Neural Network đem đi thực nghiệm trên tập test và kết quả mô hình đem lại tốt, cụ thể là 97.5%

NHÌN LẠI QUÁ TRÌNH LÀM ĐỒ ÁN

- Khó khăn:
  - Khó khăn trong việc đi tìm nguồn dữ liệu chính thống với thông tin chính xác, ít nhiễu, ít thô
  - Khó khăn trong việc tìm hiểu các siêu tham số của các mô hình được chọn để test
  - Khá gấp rút làm đồ án vì vướng trong thời gian thi cử
  - Khó khăn trong việc tìm hiểu knowledge domain của dữ liệu
  - Khó khăn trong việc tiền xử lý dữ liệu
- Những điều hữu ích học được:
  - Được làm việc hoàn chỉnh trên một mô hình khoa học dữ liệu
  - Kĩ năng khám phá và tiền xử lí dữ liệu
  - Kĩ năng đọc và tìm hiểu dữ liệu, cũng như nghiên cứu các mô hình
- Nếu có thời gian thêm, chúng em sẽ:
  - Dành thời gian nghiên cứu nhiều hơn về siêu tham số trong các mô hình, xem thử có thể tối ưu hơn nữa hay không
  - Chuẩn bị slide báo cáo hoàn chỉnh hơn, tìm hiểu github kĩ hơn
  - Tìm hiểu kỹ hơn về knowledge domain, tiền xử lí dữ liệu sạch và chuẩn hơn

TÀI LIỆU THAM KHẢO

- 1. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.24.0 documentation (scikit-learn.org)
- 2. sklearn.svm.SVR — scikit-learn 0.24.0 documentation (scikit-learn.org)
- 3. FoodData Central (usda.gov)
- 4. Composition of Foods (usda.gov)

THE END

Chúng em xin chân thành cảm ơn sự lắng nghe của thầy ạ ☺