데이터과학 기법을 적용한 맥도날드 메뉴 영양소 분석

김동준*, 임경석**
*, ** 서울 청담고등학교[1]
*e-mail : junkim100@gmail.com

Analysis on Nutrition Facts of McDonald's Menu using Data Science Techniques

Kim Dongjun*, Lim Kyungseok**
* Seoul Chungdam High School

요 약

데이터과학의 분석기법을 사용한 문제해결 방법은 많은 분야에서 활용되고 있다. 본 연구에서는 청소년들이 즐겨먹는 맥도날드 메뉴의 영양소 데이터를 분석하고 시각화를 통해 새로운 가설을 설정하고 새로운 발견을 할 수 있는 연구를 진행하였다. 영양소에 따라 건강한 메뉴와 해로운 메뉴를 구분하고자 하였으며, 데이터 분석을 통하여 새로운 건강음식 인덱스를 설정하고 그에 따른 재분석을 통하여 맥도날드 메뉴에 대한 새로운 발견을 하게 되었다.

1. 서론

파이썬 프로그래밍언어는 데이터과학 분야에서 최근 가장 많이 사용되는 도구이다[1,2]. 데이터과학을 위한 다양한 라이브러리들을 제공하고 있으며 특히 시각화 도구들이 많이 제공되어 포괄적이고 직관적인 분석을 도와준다. 본 연구에서는 파이썬 프로그래밍언어와 데이터 과학의 분석 및 시각과 기법을 이용하여 맥도날드 메뉴의 데이터를 분석하였다. 건강한 메뉴와 비건강 메뉴를 고르는 기준을 데이터 분석을 통하여 정하고 분석하고자 하였다

이를 위하여 다음과 같은 순서로 진행하였다

1) 구글의 데이터과학 공유 사이트인 케글(kaggle)에 공개된 맥도날드 메뉴 영양소 데이터를 사용하였다[3]. 이 데이터집합은 모두 25 개의 영양소 항목을 가지며, instances 의 수는 257 개이다.
2) 다운받은 데이터를 가지고 영양소들 간의 상관관계를 시각적으로 분석하여 특이점 파악하였다.
3) 특이점에 따른 새로운 가설을 세우고 그 가설을 검증할 수 있는 분석 실험을 진행하였다

Kaggle 커뮤니티에 게시된 데이터집합이라서 다양한 분석들이 그동안 시도되었는데, 대부분 파이썬의 시각적 기법을 다양하게 적용한 것이며[4], 본 논문은 세밀한 분석을 통하여 건강/비건강 메뉴를 구분하는 새로운 기준을 제시하고 분석한 결과를 보이고자 한다.

2. 영양소 데이터 분석 및 결과 도출

의학사전에 정의된 건강한 음식과 해로운 음식은 모두 해당 메뉴가 함유하고 있는 영양소의 내용과 그 영양소의 기준치(% daily value)로 정의되고 있다. 예를 들어 섬유질이 20% 이상 포함되어 있는 메뉴는 건강한 음식이고 포화지방이 20% 이상 포함된 메뉴는 건강하지 않은 음식이다. 그 정의를 우리 연구에서는 다음과 같이 다시 정의하였다.

- 건강((nutrious)성분점수 = 섬유질(%) + 비타민 A(%) + 비타민 C(%)+단백질+칼슘(%)
- 비건강(non-nutrious)성분점수 = 포화지방(%) + 나트륨(%) + 콜레스테롤(%)

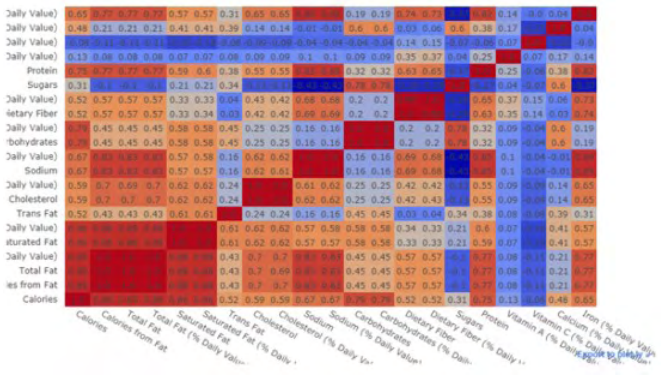
위 정의에 의해 맥도날드 메뉴를 기초 분석 하였다. 놀랍게도 그 결과에서 많은 메뉴 아이템이 건강과 비건강에 동시에 추천되는 것을 볼 수 있다. [표 2]에서 테두리로 표시한 칸은 건강점수와 비건강 점수가 동시에 높은 항목을 보여주고 있다. 예를 들면, '더블쿼터파운더'와 '빅브렉퍼스트핫케익'은 건강 아이템과 비건강 아이템에 모두 상위에 랭크되어 있는데 그 이유는 영양성분이 양쪽 모두의 것을 가지고 있기 때문이다. 따라서 이러한 기준을 사용하게 되면, 비건강 성분을 가린 채, 건강 성분 기준만 내세워 마치 건강한 메뉴인 것처럼 오해하게 될 가능성이 있다.

위와 같은 부분을 더 관찰하기 위하여, 각 영양소들간의 연관성을 파이썬에서 제공해주는 시각적 분석 도구를 사용하였다. [표 1]의 heatmap 은 각 영양소들간의 상관관계를 색상으로 보여주고 있는데, 예를 들면

[1] *청담고등학교 3 학년, **청담고등학교 교사

단백질(건강 성분)과 포화지방(비건강 성분)은 함께 나타나는 밀접한 상관성을 보여준다.

표 1 각 영양소간의 상관관계를 보여주는 히트맵



따라서 영양소를 기준으로 한 건강/비건강 메뉴 정의는 잘못된 의사결정을 유도할 수 있으므로, 우리는 본 연구에서 새로운 기준을 제안을 한다. 새로운 기준은 메뉴 아이템의 함유 영양소 성분만을 가지고 독립적으로 정의하는 것이 아니라, 두 기준 간의 균형(balance)정도를 가지고 정의하고자 하였다. 영양소 간의 비율의 차이로 새롭게 정의한 기준은 다음과 같다.

- Health_Index = (건강성분점수) - (비건강성분점수).

위의 health_index 는 건강메뉴인덱스를 의미한다. 이 점수가 0 보타 클 경우는 건강한 메뉴, 아닐 경우에는 비건강 메뉴라고 기준을 삼고, 멕도날드의 메뉴를 다시 분석을 하였다.

[표 2]에서는 새롭게 정의된 기준값을 기존의 성분 점수값들과 함께 보여주고 있다. 이 새롭게 제안된 기준에 의하면, 건강한 메뉴와 비건강한 메뉴를 분명하게 구분할 수 있으며, 이전에 독립적으로 함으로써 발생했던 잘못된 의사결정을 방지할 수 있게 되었다. 예를 들어, '빅브렉퍼스트'는 건강성분점수로만 보면 간강메뉴에 속했지만, 새로운 건강메뉴인덱스에 의하면 확실하게 비건강 메뉴로 분류된다.

표 2 각 메뉴 아이템에 대한 건강성분 점수, 비건강선분 점수, 제안한 건강메뉴인덱스 점수

Category	Item	w_nutritious	nonnutritious	w_health_index
Beef & Pork	Hamburger	53	45	8
Beef & Pork	Quarter Pounder with	173	148	25
Beef & Pork	Double Quarter Pound	168	202	-34
Breakfast	Fruit & Maple OatMea	275	15	260
Breakfast	Big Breakfast (Regular	120	337	-217
Breakfast	Big Breakfast with Hot	185	386	-201
Breakfast	Big Breakfast with Hot	168	378	-210
Chicken & Fish	Premium McWrap Chic	205	81	124
Chicken & Fish	Premium McWrap Sou	260	114	146
Chicken & Fish	Chicken McNuggets (2(	113	169	-56
Salads	Premium Southwest Sa	382	18	364
Salads	Premium Southwest Sa	450	64	386
Salads	Premium Bacon Ranch	292	88	204

3. 결론

데이터과학의 분석기법을 사용하여 메뉴에 포함된 영양소에 대해 몰랐던 패턴을 찾아낼 수 있었다. 이러한 과정을 통하여 본 연구에서는 메뉴에 포함된 영양소들의 비교 비율에 따라서 결정할 수 있도록 새로운 기준 인덱스를 제시하고 그것에 따라 분류할 수 있었다.

데이터 분석방법에서 시각화(visualization)은 직관적이고 포괄적인 관점에서의 판단을 하는데 도움을 준다. 파이썬에서 그러한 시각화 도구를 다양하게 제공해 주고 있다. 예를 들면, 비교 분석을 한눈에 할 수 있는 히트맵은 전체 패턴 분포를 파악하는데 효과적이었다.

많은 분야에서 데이터가 계속 수집이 되고 있으므로, 그러한 데이터를 효율적으로 분석하는 기법을 활용하면 그 분야에서의 새로운 패턴을 발견하게 되어 유용하게 활용할 수 있게 될 것이다.

참고문헌

[1] 드미트리 지노비에프. "모두의 데이터과학 with 파이썬", 2017
[2] Wes McKinney. "Python for Data Analysis", 2017
[3] Kaggle: Your Home for Data Science, http://www.kaggle.com
[4] N Tiwan and V Gatty. "Data Analysis on 'Nutrition Facts for McDonald's Menu' Data-set using Python, IJESC Vol.7-6, 2017