**Accurate prediction of obesity**

**-Based on body fat percentage-**

**DM Health Care Team Project**

텍스트이(가) 표시된 사진

자동 생성된 설명

2023.06.14

산업경영공학과 5조

김은태 (2019101850) 박건우 (2019100873)

박주영 (2019100877) 이명현 (2019100886)

이상재 (2019100887) 이용연 (2019100894)

발표 영상 링크: <https://drive.google.com/file/d/1CaqWXlBhwAL7UeWZfXxW77g8_qmLEa5b/view?usp=share_link>

목차

1. 비즈니스 이해 (Business Understanding)

1) 비즈니스 문제와 관련된 배경 정보

2) 프로젝트 목적 및 목표 설명

3) 프로젝트의 성공 기준

2. 데이터 이해 (Data Understanding)

1) 데이터 수집 방법 및 출처

2) 데이터 변수 및 속성 설명

3) 데이터 품질 검토 (결측치, 이상치, 불일치 등)

3. 데이터 준비 (Data Preparation)

1) 데이터 정제 및 전처리

2) Decision Tree 변수 선택 및 특성 엔지니어링

3) Regression 변수 선택 및 특성 엔지니어링

4. Modeling

1) Decision Tree

2) Regression

5. Evaluation

1) Decision Tree

2) Regression

6. Interpretation

<Report 요약문>

기존 BMI와 인바디의 제한점에 주목하여 이를 해결할 수 있는 체지방률 기반 비만 예측 모델을 만들었다. 프로젝트의 목적은 환자 주도의 데이터 공유를 통한 맞춤형 건강서비스이다. 이를 이용하여 비만으로 인한 사용자의 건강 이상 예방과 추적 관리를 목표로 하고 있다. 국민건강영양조사 원시자료를 이용하여 모델 구축에 이용했으며 비만과 상관관계가 있는 요인이 될 수 있는 변수를 선택했다. 데이터 품질 검토 후 모델에 맞게 전처리하는 과정을 거쳤다. Decision Tree와 Regression을 이용하여 비만 예측 모델을 구축했다.

1. Decision Tree   
1) 변수 선택 및 특성 엔지니어링  
BMI를 계산하여Feature Construction을 진행 후 연속형 값이 아닌 것은stratify sampling을 이용해서 값을 대체했지만, 남은 특성 또한 결측치 처리 특성과 상관계수가 0.5미만인 결과가 나와 KNN대치를 선택했다.

2) Modeling  
 2008년 데이터셋으로 정상, 과체중, 비만 (3개의 클래스)를 예측하는 모델  
Filter approach (교차 검정), Filter approach (분산 기반), Wrapper approach (Forward elimination): SFS(Sequential Feature Selector) 알고리즘, Wrapper approach (Backward elimination): RFE(Recursive Feature elimination) 알고리즘을 사용하여 모든 특성으로 모델 훈련을 진행했다. 이후 성능을 더 높이기 위해 기존 Wrapper approach에 그리드 서치를 추가한 Wrapper approach (backward) + Gridsearch, AccWrapper approach (forward) + Gridsearch를 진행했다.  
그 후 class1의 부족한 데이터 수를 증강시키기 위해 smote기법을 활용하여 Wrapper approach (backward) + Gridsearch + SMOTE를 진행했지만 데이터 증강 전과 후가 성능이 비슷했다.

2008년 + 2009년 데이터셋으로 정상, 과체중, 비만을 예측하는 모델  
추가 데이터를 확보하여 기존에 진행했던 모델링 과정을 동일하게 진행했지만 성능이 떨어졌다.

2008년 데이터셋으로 만든 모델, 2008+2009년 데이터셋으로 만든 모델  
다중 클래스 분류의 정확도를 80%로 끌어올리는 것은 불가능하다고 판단하여 정상과 비만 2개의 클래스를 예측하는 2-way split문제(정상 <-> 과체중+비만)로 BP를 수정했다. 그 결과 성능이 상승했으나 초기목표인 정확도 80%를 만족시키기 위해 데이터셋에 추가적인 전처리를 진행했다.

추가 전처리 진행 후의 모델                                                       
Nominal attribute를 제외한 특성 이상치 처리, Nominal attribute에 대한 이상치 처리를 진행했다.                                    
추가 전처리를 진행한 2008+2009년 데이터셋으로 정상, 비만 (2개의 클래스)를 예측하는 모델을 만들었지만, 성능이 오히려 급격히 줄어들었다.

이후 더 이상의 추가 분석은 불가능하다고 판단하여 최종 모델은 성능이 가장 높았던 2008년 데이터셋으로 훈련한 모델을 선택했다. 남성 모델의 성능은 Accuracy: 0.76, Precision: 0.74이다. 여성 데이터에 대해서도 동일한 과정을 통해 모델 성능을 비교했으며 여성 모델의 경우 2008년 + 2009년 데이터셋으로 만든 것의 성능이 가장 높았다. 성능은 Accuracy: 0.75, Precision: 0.67이다.

2. Regression

남성과 여성 별 Regression 모델을 만들었다. 먼저, Nominal data에 존재하는 결측치를 처리했다. 결측치 대치 모델은 scikit-learn의 RandomForestClassifier로 RandomizedSearchCV를 이용하여 RandomForest안에 들어가는 DT의 수, 최대 깊이, 최소 분할 샘플, Leaf node가 되기 위한 샘플 수 등 최적의 hyperparameter를 찾아 결측치 대치를 진행했다. Outlier 제거를 위해 DBSCAN 알고리즘을 사용했다. 하이퍼파라미터 조절을 통해 제거할 Outlier의 수를 결정했다. Outlier 제거 후, 연속형 변수와 Ordinal 변수에 대해 Min-Max scaling을 실시한 뒤 모델을 생성한 결과, Test score가 음수가 나왔다. 이를 통해 Nominal 변수에 문제가 있다고 판단하여 Nominal 변수를 제거하고 위의 전처리 과정을 거친 뒤 최종 모델링을 실시했다. 여성 모델도 전처리 과정은 같지만 여성의 경우, 여성 대상 연속형 Attribute 4가지가 추가되어 해당 Attribute에 대한 추가 전처리를 실시한 뒤 모델링을 실시했다.

Simple Linear Regression을 기본 모델로 설정했으며 독립변수 X에는 체지방률을 제외한 나머지 변수, Y에는 체지방률 변수를 할당했다. 변수 선택은 RFECV 방법을 사용했으며 RFECV는 RFE와 비슷한 방식으로 변수 중요도가 낮은 변수를 제거하며 교차 검증을 사용하여 성능을 평가한다. 성능 평가는 교차 검증 점수인 cross-validation score를 사용했다. cross-validation-score는 전체 데이터셋을 k개의 하위 집합으로 나누고, k번의 평가를 실행하여 Score를 측정한다. 이때 테스트 세트를 중복 없이 변경하면서 평가를 진행한다. 모델에서는 k=5로 설정했다.

남성 모델의 경우 총 23개의 변수가 선택되었으며 각 변수 별 계수를 확인했다. Test set에 대한 R-Squared 값은 0.67이다. 가장 유의미한 변수 상위 3개는 허리 둘레, 체중, 신장이며 선택된 변수 중 흥미로운 변수는 한약 복용 여부였다.

여성 모델의 경우 총 39개의 변수가 선택되었으며 각 변수 별 계수를 확인했다. Test set에 대한 R-Squared 값은 0.6131이다. 남성 모델과 동일하게 가장 유의미한 변수 상위 3개는 허리 둘레, 신장, 체중이였다. 흥미로운 변수로는 임의 체중감량제 섭취 여부, 카로틴 섭취량이 있었다.

1. 비즈니스 이해 (Business Understanding)

1) 비즈니스 문제와 관련된 배경 정보 ① BMI는 신장과 체중을 이용하여 비만도를 나타내는 지수이며, 구하는 방법은 이다. 다른 비만 측정법보다 간단하게 구할 수 있지만 근육량과 체지방량을 고려하지 못한다는 제한점이 있다. ② 인바디를 이용한 측정법의 경우 체지방률을 측정해 비만을 판단하기 때문에 비만 여부가 정확하다는 장점이 있다. 그러나 일반인이 체성분 기기를 이용하려면 보건소, 헬스장 등에 방문해야 한다는 번거로움이 있으며, 시중에 판매하는 기기 중 성능이 보장된 제품의 경우 22~300만 원으로 구매하기에 부담이 있다. ③ 질병관리청과 「국민건강영양조사」 통계자료에 따르면 우리나라 비만율이 2012년~ 2021년까지 남녀 통틀어 32.4%에서 37.1%로 전체적으로 증가하고 있다. 가장 눈에 띄는 변화는 2020년 남성 비만율이 41.8%에서 48.0%로 증가했다는 것이다. 그 원인을 2019년 코로나 발병과 함께 증가한 건강관리에 대한 관심으로 증가한 근력운동 열풍 때문이라고 생각했다. 이를 통해 근육량이 증가한 남성이 여성보다 더 크게 증가하여 이런 현상이 나타났다는 가설을 세웠다.

2) 프로젝트 목적 및 목표 설명 체지방률을 기반으로 한 의사결정 나무 모델을 이용하여 BMI보다 정확하게 비만을 예측한다. 특히, 마른 비만이거나 근육이 많은 사람을 대상으로도 비만 여부를 파악할 수 있게 하는 것이 핵심이다. 더 나아가 대략적인 체지방률을 예측하는 회귀 분석 모델을 만들어 사용자 주도적인 비만 관리를 돕는다. 모델 구축에 사용할 모든 변수는 사용자의 입력을 받거나, 스마트워치를 활용하여 실시간으로 습득할 수 있다. 이를 바탕으로 환자주도의 데이터 공유를 통한 맞춤형 건강서비스를 구현할 수 있다. 이는 바이오 헬스케어 분야 발전과 더불어 비만으로 인한 사용자의 건강 이상을 예방할 수 있고, 추적 관리에도 도움을 줄 수 있다.

3) 프로젝트의 성공 기준 체지방률을 기준으로 BMI를 사용하는 방법보다 실제 비만 여부를 더 정확하게 예측하고 사용자가 쉽게 자신의 건강정보를 입력할 수 있는 모델을 구축한다. 이를 토대로 사용자에게 실용적인 헬스케어 서비스를 제공한다. 비만의 기준이 되는 체지방률은 차용한 논문의 기준을 참고하였다.

2. 데이터 이해 (Data Understanding)

1) 데이터 수집 방법 및 출처 학습 데이터는 마이크로데이터 통합 서비스인 MDIS에서 찾은 국민건강영양조사 원시자료이다. 국민건강영양조사는 국민건강증진법 제16조에 근거하여 시행하는 국민의 건강행태, 만성질환 유병현황, 식품 및 영양섭취실태에 관한 법정 조사로 보건복지부 보건정책팀에서 시행하고 있다. 단, 체지방률을 조사한 연도는 2008년도가 유일하여 해당 데이터를 사용했다. 원시 데이터는 7개의 세부 DB로 나누어져 있어 데이터 통합을 진행한 후에 사용했다.

2) 데이터 변수 및 속성 설명 국민건강영양조사의 데이터는 크게 사회경제적위치지표, 건강설문조사, 검진조사, 영양조사, 여성건강으로 구성되어 있으며 각 내용 별 세부 질문에 대한 값이 포함되어 있다. 수많은 변수 중, 비만을 초래하는 요인이 아니더라도, 상관관계가 있는 요인이 될 수 있는 변수를 결정했다. 그 예시로 ‘상대적으로 소득이 적은 가구는 소득이 많은 가구보다 영양 섭취를 골고루 할 수 없기 때문에 비만의 가능성이 높다.’라는 가정이 있으며, 이때 변수는 소득분위이다.

3) 데이터 품질 검토 (결측치, 이상치, 불일치 등) 먼저 변수들을 데이터 타입별로 나누어 검토하였고, 변수의 내용 중 '모름', '무응답', '비해당'과 같은 문제가 되는 값을 발견했다. 값이 가질 수 있는 항목에 따라 평균값을 사용하거나 질문에 맞게 수정했다. 결측치의 경우에는 최빈값이나 평균값 등으로 대체했고, 약 80개의 결측치가 포함된 행은 과감히 삭제했다. 또한 결측치가 대부분이거나, 해석할 수 없는 값으로 이루어진 속성은 추후 모델링에서 제외하였다.

3. 데이터 준비 (Data Preparation)

1) 데이터 정제 및 전처리 과정 설명 국민건강영양조사 원시자료 DB에서 SAS 파일 형식인 7가지 데이터셋을 확보했으며, pandas를 이용하여 전체적인 데이터를 확인했다. 7가지 데이터셋을 ID 속성을 key로 약 1,160개 속성을 가진 하나의 데이터 시트로 통합했다. 일차적으로 필요한 속성과 행을 선별했다. 속성은 개발자의 주관적 판단으로 체지방률과 연관이 없거나, 값이 서술형인 경우를 제거했다. 서비스 측면에서 사용자가 쉽게 답하기 어려운 속성 또한 제외했다. 총 123개의 변수를 도출했다. 체지방률이 결측치인 경우와 20세 미만인 행 또한 삭제했다. 이 밖에도 공통된 문항 종류에 대한 결측치가 존재하는 경우, 그 수가 적다면 해당 행을 삭제하였다. 체지방률에 근거한 비만 기준이 성별에 따라 다르기에 남성과 여성으로 데이터를 나누었다. 추가로 남성 데이터에서는 폐경, 초경연령과 같은 속성을 제외했다. 결측치 처리는 크게 세 가지 방식으로 진행했다. 최빈값 대체, 해당 열의 행들에 대한 분포에 맞춰 값 넣기, 여러 변수에서 결측값이 나오는 행 삭제가 그것이다.

2) Decision Tree 변수 선택 및 특성 엔지니어링 BMI를 계산하여 Feature Construction을 진행했다. 남은 결측치를 포함한 컬럼은 ['N\_DIET', 'N\_INTK', 'N\_EN', 'N\_WATER', 'N\_PROT', 'N\_FAT', 'N\_CHO','L\_LN\_FAM', 'DW\_WBT\_BMD'] 였다. 이중 연속형 값이 아닌 'L\_LN\_FAM' (가족과 저녁식사 횟수), 'N\_DIET' (식이요법 여부)는 stratify sampling을 이용해서 값을 대체했다. 남은 연속형 특성들은 모두 연속형 값이므로 선형 대치를 하려고 했으나, 데이터 프레임에서 결측치 처리 특성과 상관계수가 0.5 이상인 column이 없었다. 하지만 하나도 없는 것을 확인했고, 대안으로 KNN대치를 선택했다. KNN에서의 최적의 k값 (근처 이웃 수)을 교차검정을 통해 찾았을 때의 값은 3이었지만, 모델링 과정에서 k=5로 만든 모델의 성능이 더 높아 k=5를 사용했다.

3) Regression 변수 선택 및 특성 엔지니어링

(1) 남성  
① 결측치 처리 Nominal data에 존재하는 결측치를 처리했다. 결측치는 공통적으로 NaN으로 표현되고 Attribute별로 다른 표현(ex.9, 99)으로 되어 있는 것도 있다. Nominal attribute별로 결측치를 csv파일로 정리하여 결측치 대치를 진행했다.   
결측치 대치 모델은 scikit-learn의 RandomForestClassifier로 RandomizedSearchCV를 이용하여 RandomForest안에 들어가는 DT의 수, 최대 깊이, 최소 분할 샘플, Leaf node가 되기 위한 샘플 수 등 최적의 hyperparameter를 찾아 결측치 대치를 진행했다.    
② One-hot encoding Nominal 데이터를 사용하기 위해 One-hot encoding을 실시했다. get\_dummies함수를 사용하여 Nominal attributes를 변환한 뒤, Regression modeling에 사용했다.   
 이상치 제거 One-hot encoding 이후, Simple linear regression을 적용한 결과, test score 값이 음수가 나왔다. 이는 linear regression 특성 상 outlier가 매우 큰 영향을 미쳤다고 판단하여 outlier를 제거했다. Outlier 제거 알고리즘은 DBSCAN 알고리즘을 사용하였다. DBSCAN 알고리즘은 밀도가 높은 부분을 하나의 cluster로 판단하는 방식이다. Hyperparameter로 반경을 나타내는 epsilon, cluster로 판단하는 최소 조건인 min\_samples, 거리를 나타내는 방식인 metric가 있다. 그 결과, core points, border point, noise point가 나온다. 이 중, noise point가 outlier에 해당한다. 휴리스틱으로 접근한 결과, epsilon=0.02, min\_samples=7까지 진행했다. 차원이 매우 크기에, metric은 cosine distance로 결정했다.   
그 결과, outlier sample은 총 41개로, 남겨진 sample은 1392개였다.

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명텍스트, 라인, 도표, 스크린샷이(가) 표시된 사진

자동 생성된 설명  
 min max scaling  
데이터의 이상치 제거 후, 연속형 변수와 Ordinal 변수에 대해 MinMaxScaler를 사용했다. 하지만, 5-fold CV를 통하여 Model selection을 진행하였을 때 음수인 값이 나오고, test score 또한 다음과 같은 음수 값이 나왔다. Training score: 0.6698, Test score: -3.0131227697629804e+18  
이 현상은 outlier를 여러 번 제거한 후에도 변하지 않아 Nominal attribute에 문제가 있다고 생각했다. Nominal 변수를 제거하고 다시 전처리를 진행했다. 휴리스틱 접근법을 사용하여 이상치 탐지를 min\_samples=17, epsilon=0.015로 결정하여 DBSCAN을 진행하였다.

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명  
제거된 outlier는 114개, 남은 record는 1319개이다.   
모든 column(ordinal과 numerical)에 min-max scaling을 진행했다.  
 PCA PCA 분석의 목적은 원본 데이터의 변동성을 최대한 보존하면서 데이터를 효율적으로 표현하는 것이다. 지표로는 누적 설명된 분산 비율을 확인하여 주성분의 개수를 정했다. 99%의 누적 설명된 분산 비율을 설정하여 91개의 주성분을 사용했다. 이렇게 얻은 91개의 새로운 변수를 사용하여 linear regression 모델을 적용했더니 성능이 좋지 않아 RFECV 방법을 사용했다.  
상관관계 분석 데이터셋의 Attribute가 많아 전체를 이용하는 것보다 종속 변수와 상관관계가 높은 독립 변수를 사용하여 모델링을 진행해 보기 위해 상관관계 분석을 진행하여 상관관계가 높은 독립 변수를 선택하여 모델링을 했다. 다중공선성 문제를 해결하기 위해 독립 변수 사이의 상관관계 분석을 실시했다.

(2) 여성 남성과 전처리 과정은 같다. 하지만, 여성의 경우, 폐경 연령/초경 연령/여성호르몬제 복용기간/경구피임약 복용기간 4가지 연속형 Attribute가 추가되어 해당 Attribute의 결측치를 추가적으로 처리했다. 비해당에 대해선 0으로 값을 부여하고 무응답에 대해선 해당 Attribute의 평균값을 부여했다. 그 결과 record=1997, columns=147개이다. 이후 이상치 처리 단계는 min\_samples = 28, epsilon = 0.019로 DBSCAN을 적용했다. 제거된 sample은 219개, 잔여 record는 1780개이다.

4. Modeling

1) Decision Tree (최초 성능 목표는 Accuracy 80%로 설정했다.)  
(1) 2008년 데이터셋으로 정상, 과체중, 비만 (3개의 클래스)를 예측하는 모델 ① 모든 특성으로 모델 훈련 -> 123개(전체 개수) accuracy: 0.59 ②-1 Filter approach (교차 검정) -> 119개 Accuracy: 0.47 ②-2 filter approach (분산 기반) -> 95개 Accuracy: 0.57 ③-1 Wrapper approach (forward elimination): SFS(Sequential Feature Selector) 알고리즘 사용 -> 특성 61개 Accuracy: 0.57 ③-2 Wrapper approach (backward elimination): RFE(Recursive Feature elimination) 알고리즘 사용 -> 특성 1개 Accuracy: 0.63 \* 성능을 더 높이기 위해 가장 성능이 좋을 것으로 예상되는 기존 Wrapper approach에 그리드 서치를 추가로 진행했다. (의사결정나무 파라미터 튜닝) 이때 특성 개수도 10개로 제한했다. (남자 데이터 수인 1440개 ()를 고려하여 결정한 특성 개수임) ④-1 Wrapper approach (backward) + Gridsearch -> 10개, Accuracy: 0.68 ④-2 Wrapper approach (forward) + Gridsearch -> 10개, Accuracy: 0.60 최초 목표였던 80%에 못 미치는 결과가 도출된 원인을 분석한 결과, 다중 클래스 분류문제에서 현재 class 1에 해당하는 과체중의 범위가 정상(0), 비만(2)의 범위에 비해 좁기 때문에 class 1에 해당하는 object가 많이 부족하고, 따라서 정확도가 예상보다 낮은 것이라고 예상했다. ⑤ Wrapper approach (backward) + Gridsearch + SMOTE (Accuracy: 0.68) 따라서 class1의 부족한 데이터 수를 증강시키기 위해 smote기법을 활용하여 train set에서의 각 target class 비율을 동일하게 설정했다. (training set에서 class 0, 1 ,2 중 최대 object수인 607개로 설정) 하지만 데이터 증강 전과 후가 성능이 비슷했다.

(2) 2008년+2009년 데이터셋으로 정상, 과체중, 비만을 예측하는 모델 (Accuracy: 0.63) 데이터셋의 크기가 작기 때문에 다중 클래스 분류의 정확성이 떨어지는 것이라고 예상했다. 추가 데이터를 확보하기 위해 국민 건강 영양조사데이터 중 2009년 데이터를 사용했다. 해당 데이터를 추가한 2008+2009년 데이터 셋을 만들고, 기존에 진행했던 모델링 과정을 동일하게 진행했다. 하지만 오히려 성능이 더 떨어지는 결과가 나왔다.

(3) 2008년 데이터셋, 2008년+ 2009년 데이터셋으로 정상, 비만 (2개의 클래스)를 예측하는 모델 다중 클래스 분류의 정확도를 80%로 끌어올리는 것은 불가능하다고 판단하여, 정상, 과체중, 비만 3개의 클래스를 예측하는 것에서 정상과 비만 2개의 클래스를 예측하는 2-way split문제(정상 <-> 과체중+비만)로 BP를 수정했다. (과체중을 비만에 포함시킴) 앞서 만든 최종 모델과 동일한 알고리즘을 적용하여 2008년 데이터셋과 2008+2009년 데이터 셋 2개로 각각 모델을 학습시켜 정확도를 구했다.  
①2008년 데이터셋으로 만든 모델 Accuracy: 0.76, Precision: 0.74②2008+2009년 데이터셋으로 만든 Accuracy: 0.75, Precision: 0.73 성능이 상승했으나 초기 목표였던 정확도 80%를 만족시키기 위해 데이터셋에 추가적인 전처리를 진행하여 목표 성능을 달성하기로 결정했다.

(4) 추가 전처리 진행과 그 후의 모델 ① nominal attribute을 제외한 특성 이상치 처리 Z 스코어를 이용하여 이상치가 있다고 판단되는 특성 중 nominal column을 제외한 특성에 대해 이상치 처리를 진행했다. (데이터 유실을 최소화와 이상치를 줄이기 위해서 z는 3으로 설정함) ② nominal attribute에 대한 이상치 처리 답변 중 모름 혹은 비해당 값을 분석가 판단 하에 최빈값 또는 비율에 맞춘 값으로 대체함. - 모름 or 비해당 값의 개수가 한자리 수 -> 최빈값 대체 - 모름 or 비해당 값의 개수가 두 자리 수 -> 해당 특성 value의 비율대로 대체 - 모름 or 비해당 값의 개수가 세 자리 수 -> 해당 특성 제거 ③ 추가 전처리를 진행한 2008+2009년 데이터셋으로 정상, 비만 (2개의 클래스)를 예측하는 모델 성능이 오히려 급격히 줄어들었음. (Accuracy: 0.68, Precision: 0.67) 예상되는 원인으로는 부족한 도메인 지식으로 인한 잘못된 데이터 전처리라고 판단했고, 시간 관계상 더 이상의 추가 분석은 불가능하다고 판단했다. 따라서 최종 모델은 성능이 가장 높았던 2008년 데이터셋으로 훈련한 모델을 선택했다. **\* (3)-① 모델 (Accuracy: 0.76, Precision: 0.74)** 이때 선택된 특성은 ['age', 'N\_INTK', 'N\_EN', 'N\_WATER', 'N\_PROT', 'N\_FAT', 'N\_CHO', 'HE\_ht', 'HE\_wc', 'BMI']이었고, Tree를 시각화 했을 때 모든 특성이 사용된 것을 확인할 수 있었다. 여자 데이터에 대해서도 동일하게 2008년 데이터셋과, 2008+2009년 데이터셋, 각각의 데이터셋에 추가 전처리를 진행한 데이터셋으로 모델을 학습시켜 가장 성능이 높은 모델을 선택했다. ① 여자 2008년 데이터셋으로 만든 모델 Accuracy: 0.73, Precision: 0.61 ② 여자 2008+2009년 데이터셋으로 만든 모델 Accuracy: 0.75, Precision: 0.67 ③ 여자 2008+2009년 데이터셋 (추가 전처리 완료)으로 만든 모델 Accuracy: 0.72, Precision: 0.49 -> ②번 모델 선택, 이때 선택된 특성은 ['age', 'icm5’, 'N\_INTK', 'N\_PROT', 'N\_FAT', 'N\_CHO', 'HE\_ht', 'HE\_wc', 'HE\_wt', 'BMI']이었고, Tree를 시각화 했을 때 실제로 사용한 특성은 ['age', 'HE\_ht', 'HE\_wt']이 제외된 ['icm5’, 'N\_INTK', N\_PROT', 'N\_FAT', 'N\_CHO', 'HE\_wc', 'BMI']이었다.

2) Regression

(1) 남성  
전처리를 통해 선택된 Attribute를 가지고 독립변수 X에고는 체지방률을 제외한 나머지 변수, y에는 체지방률 변수를 할당했다. Simple Linear Regression을 기본 모델로 설정했다. 변수 선택은 RFECV 방법을 사용했다. RFE는 원하는 개수의 변수가 남을 때까지 학습을 반복하여 의미 없는 변수를 제거하는 backward 방식이다. RFECV는 RFE와 비슷한 방식으로 변수 중요도가 낮은 변수를 제거하며 교차 검증을 사용하여 성능을 평가한다. 성능이 제일 높은 변수의 개수를 최적의 개수로 선택한다. 최적의 변수 개수는 23개로 선택되었다. 성능 평가는 교차 검증 점수인 cross-validation score를 사용했다. cross-validation-score는 전체 데이터셋을 k개의 하위 집합으로 나누고, k번의 평가를 실행하여 Score를 측정한다. 이때 테스트 세트를 중복 없이 변경하면서 평가를 진행한다. 모델에서는 k=5로 설정했다.  
선택된 변수는 다음과 같다.  
['DJ4\_dg', 'DJ4\_3', 'DC2\_dg', 'LQ\_2EQL', 'BO3\_12', 'BO3\_07', 'BO3\_09', 'N\_INTK', 'N\_PROT', 'N\_FAT', 'N\_CHO', 'N\_PHOS', 'N\_NA', 'N\_VA', 'N\_CAROT', 'N\_RETIN', 'N\_B1', 'N\_NIAC', 'N\_VITC', 'HE\_ht', 'HE\_wt', 'HE\_wc', 'DW\_WBT\_BMD']

(2) 여성  
여자 데이터 모델링은 남자 데이터 모델링 방식과 동일하다. 가장 높은 성능을 가지는 Feature의 개수는 39개였다. 선택된 39개의 Feature는 아래와 같다.

['edu', 'allownc', 'DI3\_dg', 'DN1\_dg', 'DC3\_dg', 'DC4\_dg', 'DK4\_dg', 'EC\_env\_5', 'BO1\_1', 'BO1\_3', 'BO3\_01', 'BO3\_02', 'BO3\_05', 'BO3\_04', 'BO3\_12', 'BO3\_07', 'BO3\_09', 'BS5\_21', 'BS5\_22', 'HW\_ES\_DR', 'HW\_MS\_A', 'N\_WAT\_C', 'N\_EN', 'N\_PROT', 'N\_CHO', 'N\_FIBER', 'N\_ASH', 'N\_PHOS', 'N\_NA', 'N\_VA', 'N\_CAROT', 'N\_RETIN', 'N\_B1', 'N\_VITC', 'LK\_GD\_P6', 'HE\_ht', 'HE\_wt', 'HE\_wc', 'DW\_WBT\_BMD'] 선택된 Feature을 가지고 선형 회귀 모델을 학습시켰다.

5. Evaluation

1) Decision Tree

① 남자: Accuracy= 0.76, Precision: 0.74  
② 여자: Accuracy = 0.75, Precision = 0.67  
Business Problem에 따르면 비만인 사람을 정확히 예측하는 것이 중요하므로, Accuracy와 precision을 성능 평가 척도로 사용했다.  
남자 의사결정 모델 정확도의 경우 남자 데이터셋에서 확인한 BMI의 측정 정확도인 0.67보다 약 9%정도 더 좋은 성능을 보였다.  
여자 의사결정 모델 정확도의 경우 여자 데이터셋에서 확인한 BMI의 측정 정확도인 0.59보다 약 16%정도 더 좋은 성능을 보였다.

2) Regression

(1) 남성  
test set에 대한 R-Squared 값은 0.67이다. 가장 유의미한 변수 상위 3개는 HE\_wc(허리둘레), HE\_wt(체중), HE\_ht(신장)이며, 흥미로운 변수는 한약 복용 여부였다. Raw 데이터의 결측치 추정 후 대체, Numerical과 Ordinal 변수 선택, Outlier 제거, Min-Max Scaling의 과정을 거친 뒤 RFECV를 통해 가장 성능이 좋았을 때의 Feature를 선택했다. 이 Feature들을 이용하여 Simple Linear Regression 모델링을 진행했을 때 Test Set에서 가장 높은 성능이 나왔다.

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

(2) 여성  
모델링 결과, test score는 0.6131이며 R-Squared 결정 계수는 0.61이다. 가장 유의미한 변수 상위 3개는 He\_wc(허리둘레), He\_ht(신장), He\_wt(체중)이었으며 선택된 변수 중 흥미로운 변수는 BO3\_05(임의 체중감량제 섭취 여부), N\_CAROT(카로틴 섭취량)이 있었다. Raw 데이터의 결측치 추정 후 대체, numerical과 ordinal 변수 선택, outlier 제거, Min-max Scaling의 과정을 거친 뒤 RFECV를 통해 가장 성능이 좋았을 때의 Feature를 선택했다. 이 Feature들을 이용하여 Simple Linear Regression 모델링을 진행했을 때 Test Set에서 가장 높은 성능이 나왔다.

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

6. Interpretation

1) Decision Tree  
{소득, 식사량, 일일 단백질 섭취량, 일일 지방 섭취량, 일일 탄수화물 섭취량, 몸무게, BMI}의 정보를 통해 비만 예측 모델을 완성했다. 최종 선택된 Decision Tree 모델을 통해 Information gain이 가장 높은 허리둘레가 비만에 가장 큰 요인이라는 사실을 알 수 있다. 또한 BMI기준으로 비만이더라도, 다른 요인들을 거쳤을 때 비만이 아니라는 결과가 도출되는 경우도 확인할 수 있었다. 일반적으로 비만을 판단하는 가장 간단한 지표인 BMI로만 실제 비만(체지방률 기준)을 판단했을 때의 정확도는 약 66%이다. 우리의 모델을 활용하면 일반적으로 사용하는 BMI에 비해 약 9~16% 높은 정확도로 비만을 판단할 수 있다. 이는 원래 BP인 인바디보다 간편하고 BMI보다 더 정확한 모델을 만드는 것을 달성했다고 볼 수 있다. 향후 웨어러블 디바이스가 더 발전하면, 인바디보다 간편하고 BMI보다 정확한 서비스를 기대할 수 있다.

2) Regression  
남자의 경우 유의미한 변수 상위 10개는 {허리둘레, 체중, 신장, 다이어트 목적 한약 복용, 인 섭취량, 골밀도, 원푸드다이어트로 인한 체중 조절, 비타민C 섭취량, 티아민 섭취량, 탄수화물 섭취량}이었다.  
여자의 경우 유의미한 변수 상위 10개는 {체중, 키, 허리둘레, 체중감량제 섭취, 카로틴 섭취량, 의지로 금연, 니코틴 대체용품으로 금연, 다이어트 목적 한약 복용, 여성호르몬제 총 복용기간, 원푸드다이어트로 인한 체중조절}이었다.   
위와 같은 결과를 바탕으로 남녀 모두 허리둘레, 체중, 키가 제일 중요한 변수라는 사실을 확인했고, 영양소 섭취가 중요하다는 사실을 알 수 있었다. 체중 조절과 관련된 변수가 중요한 이유는 체중 조절에 신경쓰고 노력하는 것 자체가 체지방률과 연관이 있다는 것으로 해석했다.  
남자에서 보이는 특별한 변수는 골밀도가 있는데, 체중에 뼈 무게 또한 포함되어 있기에 중요한 변수라고 해석하였다. 여자에서 보이는 특별한 변수는 여성호르몬제 총 복용기간이 존재하였는데, 여성호르몬제를 복용하면 살이 찐다는 부작용을 보여준다고 해석했다.

참고문헌 (분량 제한 없음)

1. 김지현. 매일경제. 2016.04.27. BMI(체질량지수)는 정말 ‘건강’을 대변해줄까. <https://www.mk.co.kr/news/society/7320153>

2. 성순창, 김현수. (2013년). 한국성인의 체질량지수와 제지방량지수 및 체지방량지수와의 관련성. 운동과학(서울과학기술대). 22(4). 301-308

3. 보건복지부 질병관리본부. 2008.12. 국민건강영양조사 원시자료 이용지침서

4. 전종보. 헬스조선. 2022.10.02. 마른 비만 더 무서워… 체지방률 ‘00%’ 넘으면 진단. <https://health.chosun.com/site/data/html_dir/2022/09/30/2022093002281.html>

5. 정운경. 하이닥. 2019.05.25. BMI말고 체지방률로 스마트하게 내 몸을 관리하라. <https://www.hidoc.co.kr/healthstory/news/C0000462861>

6. 정윤식. 의학신문. 2019.01.03. 마른비만? 겉보기 괜찮아도 건강 적신호! <https://health.chosun.com/site/data/html_dir/2022/09/30/2022093002281.html>

7. 질병관리청. 2023.03.03. 최근 10년 간 성인, 중고등학생 모두 비만 유병률 증가. <https://www.korea.kr/news/policyNewsView.do?newsId=148912348>

8. Goonasegaran AR · Nabila FN · Shuhada NS. (2012). Comparison of the effectiveness of body mass index and body fat percentage in defining body composition. Singapore Medical Journal, 53(6)

개인별 기여 사항 (1인 구성팀은 아래 본인의 기여 사항만 적으면 됩니다.)

| 성명 | 기여 내용 (기여내용은 개인당 최대 1페이지 분량으로 작성) |
| --- | --- |
| 김은태 | Regression 중 RandomForestClassifier를 통한 결측치 대체.  Regression 중 DBSCAN을 이용한 heuristic 기반 Outlier 제거.  model selection 중 5-fold CV 도입.  matplotlib을 통한 y\_true, y\_predicted의 관계 plotting.  코드 정리  전처리, interpretation 중 regression 파트 보고서, PPT 및 대본 작성. |

| 성명 | 기여 내용 (기여내용은 개인당 최대 1페이지 분량으로 작성) |
| --- | --- |
| 박건우 | 건강생활습관 데이터를 살펴보고 건강생활습관 데이터를 통해 비만 여부를 예측하는 분류모델을 만들자는 idea를 제시했습니다. 비즈니스 이해에 대한 초안을 작성하여 팀원들에게 제공했습니다. 이후 체지방률을 예측하는 Regression팀과 비만 여부를 판단하는 Decision tree로 나뉘었을 때 Decision tree팀원으로 활동했습니다. 프로젝트 진행 전 사전조사를 통해 데이터 분할에 대한 기본지식 및 방향을 제시했습니다. (설명변수, 목표변수, stratify, validation set을 따로 분할하지 않고 k-fold로 성능평가 등 )  저는 DT팀에서 데이터 전처리 및 모델링을 담당했습니다. 전처리 과정에서 연속형 특성의 결측치를 처리하기 위해 상관계수를 이용한 선형 대치를 시도했으나, 각 연속형 특성과 상관계수가 0.5 이상인 특성이 없다는 것을 확인하고 대안으로 KNN 알고리즘을 사용하여 결측치를 대치했습니다. 처음에는 임의로 k=5를 사용하여 결측치를 처리했으나, 교차검정을 통해 최적의 k값을 알아 낼 수 있다는 것을 알고 k=3으로도 결측치를 처리했습니다.  결측치 처리를 마친후에는 수업시간에 다뤘던 feature selection방법 및 코드를 검색하여 전체 특성, filter approach(교차검정, 분산기반), wrapper appraoch(forward, backward)를 각각 실행하는 코드를 작성했습니다. 이를 통해 wrapper approach가 대부분의 경우에서 더 좋은 성능을 보이는 것을 확인하고, wrapper approach 코드에는 그리드 서치를 통해 하이퍼 파라미터 튜닝까지 진행했습니다. 하지만 기대 이하의 성능이 나오는 것을 확인하고 그 원인으로 다중 클래스 분류 문제에서 과체중 class(1)의 object의 수의 부족이라고 생각했습니다. 이를 해결하기 위한 방법으로 SMOTE를 적용하여 train set의 과체중 object 수를 증강시켜 모델을 학습시켰습니다. 하지만 증강하기 전과 후의 성능이 비슷하다는 것을 확인했고,  다시 성능을 개선시키기 위해 자료를 수집했던 MDIS에서 체지방률과 기본 DB가 따로 저장되어있던 2009년, 2010년 자료를 추가로 확보했습니다. 기본 DB와 체지방률 DB를 공통 ID를 기준으로 데이터를 합친 후 2008년 자료를 전처리했던 방식과 동일하게 진행하여 2008년+2009년+2010년 데이터(남/여 구분)를 만들고 모델링을 재진행했습니다. 하지만 성능이 개선되지 않아 정상과 (과체중, 비만)으로 이진분류를 해보자는 의견을 제시했습니다. 그러자 성능이 76%까지 증가함을 확인하고 최종 모델로 선택하게되었습니다. 추가적으로 남/여를 구분하지 않은 모델, 2008+2009+2010 총 3년간의 데이터를 병합한 데이터셋으로 만든 모델, SMOTE를 사용한 모델과 사용하지 않은 모델, 2008년 데이터셋의 과체중 object만 뽑아서 2009년 데이터셋에 추가하여 만든 모델 등 다양한 모델을 만들어서 성능을 비교해보았습니다. 다양한 데이터셋으로 모델링을 진행하여 f1score를 확인했을 때 과체중 class에 대한 예측이 증가하긴하지만, 그만큼 다른 클래스의 정확도가 떨어지는 것을 발견하여 과체중의 범위가 다른 클래스에 비해 너무 좁아서 현재 상황으로는 다중 클래스 예측은 한계가 있다고 결론짓고 이진분류를 선택했습니다. |

| 성명 | 기여 내용(기여내용은 개인당 최대 1페이지 분량으로 작성) |
| --- | --- |
| 박주영 | 변수선택 기법 중 하나인 PCA 분석을 진행하여 성능을 확인함  regression 파트 데이터 변수 정리  regression 파트 데이터 전처리 min-max scaling  regression 파트 남성 모델링 부분 보고서 작성  regression 파트 남성 모델링 부분 ppt  regression 파트 녹화 |

| 성명 | 기여 내용(기여내용은 개인당 최대 1페이지 분량으로 작성) |
| --- | --- |
| 이명현 |  |

| 성명 | 기여 내용(기여내용은 개인당 최대 1페이지 분량으로 작성) |
| --- | --- |
| 이상재 | 의사결정나무 모델에 적합한 데이터 전처리중, nominal attribute 전처리 진행  의사결정나무의 성능을 개선하기 위해 하이퍼파라미터 튜닝, 그리드 서치 진행(max\_depth, min\_sample\_split, min\_sample\_node, criterion(gini,entropy))  의사결정나무 모델의 시각화 진행 (matplotlib과 sklearn의 tree 사용)  의사결정나무 정확도 향상을 위해 추가 이상치 처리 진행(자체적인 기준에 맞춰 nominal, ordinal data의 분산 줄임, 모름이나 비해당이라고 답이 되어있는 설문을 최대한 제거  모름 or 비해당 값 한자리수 -> 최빈값 대체 , 모름 or 비해당 값 두자리수 -> 해당 특성 value의 비율대로 대체, 모름 or 비해당 값 세자리수 -> 해당 특성은 오히려 분석 의도를 흐린다고 판단하여 특성을 제거 |

| 성명 | 기여 내용(기여내용은 개인당 최대 1페이지 분량으로 작성) |
| --- | --- |
| 이용연 | 의사결정나무 모델에 사용할 데이터 전처리  연속형 데이터 범주형 데이터로 변환  selectKbest를 통한 특성 추출 후 의사결정 나무 모델 구축, 성능 개선 (하이퍼 파라미터 튜닝)  의사결정나무 모델 시각화  보고서 요약  ppt 작성 |

팀 자체 평가 (팀원이 모두 논의해서 1페이지로 작성)

| **평가 항목** (괄호 안의 숫자는 점수) | **점수** | **관련 보고서 내용**  기재 예시: P.2, L.3-4 |
| --- | --- | --- |
| ○ Proposal에 기재된 목적, 목표, 추진전략, 절차의 달성도 (10) | 9 | P.11, L.11-18, L21, L28  P12, L.9-17, L.19-24 |
| ○ CRISP-DM Business Understanding 단계의 수행 충실성 (10)    - 해당 단계의 프로젝트 수행 결과가 구체적이며 충실함    - 해당 단계의 프로젝트 추진방식, 절차 등이 타당함. | 10 | P.5, L.1, L.28 P.6, L.13  P.8, L.23 P.10, L.20  P.11, L.9 P.12, L.7 |
| ○ CRISP-DM Data Understanding 단계의 수행 충실성 (10)    - 해당 단계의 프로젝트 수행 결과가 구체적이며 충실함    - 해당 단계의 프로젝트 추진방식, 절차 등이 타당함. | 10 | P.5, L.30-34  P.6, L.8-12 |
| ○ CRISP-DM Data Preparation 단계의 수행 충실성 (20)    - 해당 단계의 프로젝트 수행 결과가 구체적이며 충실함    - 해당 단계의 프로젝트 추진방식, 절차 등이 타당함. | 20 | P.6 15-24, L26-29  P.7, L.7-9, L.16-22  P.8, L.1-5, L.17-22 |
| ○ CRISP-DM Modeling 단계의 수행 충실성 (20)    - 해당 단계의 프로젝트 수행 결과가 구체적이며 충실함    - 해당 단계의 프로젝트 추진방식, 절차 등이 타당함. | 19 | p.8, L23 - p.10,L.19  P.10, L.22-34,  P.11 L.1-8 |
| ○ CRISP-DM Evaluation 단계의 수행 충실성 (10)    - 해당 단계의 프로젝트 수행 결과가 구체적이며 충실함    - 해당 단계의 프로젝트 추진방식, 절차 등이 타당함. | 10 | p.11, L.10-L.18  P.11, L.21-28,  P.12, L.1-6 |
| ○ CRISP-DM Interpretation 단계의 수행 충실성 (15)    - 해당 단계의 프로젝트 수행 결과가 구체적이며 충실함    - 해당 단계의 프로젝트 추진방식, 절차 등이 타당함. | 15 | p.12, L.7-11  p.12, L.17-29 |
| ○ 프로젝트 결과의 가치 창출 가능성 및 파급효과 (5)  - 프로젝트 결과를 통한 기술·경제·사회적 가치 창출 가능성  - 프로젝트개발 성과물의 활용 및 확산 효과 | 5 | p.12, L. 12-16 |

\* 3번째 column ‘관련 보고서 내용’ 에는 평가 점수의 근거가 되는 제안서 내용의 위치를 기재하세요. 기재방법의 예시인 P.2, L.3-4가 의미하는 바는 ‘제2페이지의 3번째에서 4번째 줄에 관련 내용이 있음’입니다.