1. **연구 목적**

본 연구의 목적은 총 106개의 문항과1,497명의 응답을 토대로 ‘운동/다이어트앱 이용 여부’를 예측모델의 구현을 위해, 어떤 문항이 영향력((Influential factor)의 우선순위가 높은지를 찾아내는데 있다.

**2. The results of priority ordering of the Influential factor (=feature importance)**

* **총 106개 factors - 상위 20개 정리**

건강생활을 위한 앱 이용 : 12%,   
주관적규범\_2 : 5%,   
주관적규범\_3 : 4%,   
소셜미디어이용빈도: 2%,   
스마트폰앱이용갯수: 2%,   
주관적규범\_1 : 2%,   
체중조절 : 2%,   
다이어트의향\_7 : 2%,  
운동에대한태도\_5 : 2%,   
운동의유익성\_5 : 2%

**2. Priority analysis method**

Preparation) 분석을 위해 train set과 test set은 7:3(1197개 300개)의 비율로 랜덤하게 나누었으며, 각각 target(운동/다이어트앱 이용) class의 분포는 6:4로 동일하게 나누었다.

**2-1) First step: Machine learning model selection**

There are several machine learning models such as RandomForest, LightGBM, Adaboost XGBoost which have their own advantages and disadvantages. In order to choose the suitable leaning model, we refer to the Kaggle[[1]](#footnote-1) site. There were 29 challenges at Kaggle[[2]](#footnote-2) in 2015, among them XGBoost was used for core learning model for the 17 winning solutions. The main problems in the 17 winning solutions are related to customer behavior prediction, store sales prediction which is identical with our purpose of this study [1]. Furthermore, since our target class consists of Use and Non-Use of Fitness and Diet App, it can be solved by XGBoostClassifier. In this sense, XGBoost is adopted in this study.

**Second step: Hyperparameter tuning by GridSearchCV**

To avoid overfitting[[3]](#footnote-3) as well as optimize the hyperparameter of XGBoost, cross-validation (CV) with GridSearch is proceeded based on 10 folds in [3]. In more detail, the CV first randomly divide the trained data set into subset of k (i.e., 10) folds and then calculate prediction accuracy. Finally, the prediction accuracy is compared with that of the test set: if the accuracy is similar each other, we consider there is no overfitting in this XGBoost. During cross-validation, the hyperparameters of Gridsearch is optimized by substituting all possible numbers with brute force method.

To satisfy the main purpose to **prioritize the Influential factors**, Gridsearch is employed because the accuracy is a crucial factor. As a result, the mean accuracy score in Gridsearch is 0.754386 with the std accuracy score of 0.020116.

------------------------------------Parameter첨부 예정--------------------------------------

**Third step : Find Influential factors**

In this step, we prioritize the influential factors by using tuned XGBoost Model. Information gain method in [4] is used to identify the precedence of the influential factors among 106 factors. Table 1 reveals the top 10 influential factors to identify the primary influenced factors for the Personal and Social Predictors of Use and Non-Use of Fitness and Diet App.

Table I

건강생활을 위한 앱 이용 : 12%,   
주관적규범\_2 : 5%,   
주관적규범\_3 : 4%,   
소셜미디어이용빈도: 2%,   
스마트폰앱이용갯수: 2%,   
주관적규범\_1 : 2%,   
체중조절 : 2%,   
다이어트의향\_7 : 2%,  
운동에대한태도\_5 : 2%,   
운동의유익성\_5 : 2%

[1] XGBoost: A Scalable Tree Boosting System - Tianqi Chen

[2] Random Search for Hyper-Parameter Optimization -James Bergstra

[3] A Survey of cross-validation procedures for model selection (DOI: 10.1214/09-SS054) – Sylvain Arlot

[4] information gain

1. Data competition site : www.kaggle.com [↑](#footnote-ref-1)
2. Data competition site : www.kaggle.com [↑](#footnote-ref-2)
3. Test set에서만 잘 작동하는 현상 [↑](#footnote-ref-3)