Journal of Convergence for Information Technology
Vol. 11. No. 6, pp. 49-54, 2021
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2021.11.06.049

불균형 클래스에서 AutoML 기반 분류 모델의 성능 향상을 위한 데이터 처리

이동준[1], 강지수[2], 정경용[3*]
[1]경기대학교 AI컴퓨터공학부 학부생, [2]경기대학교 컴퓨터과학과 석사과정,
[3]경기대학교 AI컴퓨터공학부 교수

Data Processing of AutoML-based Classification Models for Improving Performance in Unbalanced Classes

Dong-Joon Lee[1], Ji-Soo Kang[2], Kyungyong Chung[3*]
[1]Student, Division of AI Computer Science and Engineering, Kyonggi University
[2]Student, Department of Computer Science, Kyonggi University
[3]Professor, Division of AI Computer Science and Engineering, Kyonggi University

요 약 최근 스마트 헬스케어 기술의 발전에 따라 일상적인 질환에 대한 관심이 증가하고 있다. 이에 따라 헬스케어 데이터를 통해 예측 모델로 질병을 분석하거나 예측하는 연구들이 증가하고 있다. 그러나 헬스케어 데이터에는 양성 데이터와 음성 데이터의 불균형이 존재한다. 이는 특정 질환을 가진 환자에 비하여 상대적으로 환자가 아닌 사람이 많아 데이터 수집에 어려움이 있어 발생하는 현상이다. 데이터 불균형은 질병 예측 및 탐지 시 진행하는 모델의 성능에 영향을 끼치기 때문에 이를 제거할 필요가 있다. 따라서 본 연구에서는 오버샘플링과 결측값 대치를 통해서 데이터 불균형을 해소한다. AutoML을 기반으로 여러 모델의 성능을 파악하고 모델 중 상위 3개의 모델을 앙상블한다.

주제어 : 데이터 불균형, 오버샘플링, 헬스케어, AutoML, 결측값 대치

Abstract With the recent development of smart healthcare technology, interest in daily diseases is increasing. However, healthcare data has an imbalance between positive and negative data. This is caused by the difficulty of collecting data because there are relatively many people who are not patients compared to patients with certain diseases. Data imbalances need to be adjusted because they affect performance in ongoing learning during disease prediction and analysis. Therefore, in this paper, We replace missing values through multiple imputation in detection models to determine whether they are prevalent or not, and resolve data imbalances through over-sampling. Based on AutoML using preprocessed data, We generate several models and select top 3 models to generate ensemble models.

Key Words : Data Imbalance, Oversampling, Healthcare, AutoML, Data Imputation

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2019R1F1A1058651).

*Corresponding Author : Kyungyong Chung(dragonhci@hanmail.net)

Received May 17, 2021
Revised June 6, 2021
Accepted June 20, 2021
Published June 28, 2021

1. 서론

최근 헬스케어 빅데이터 산업은 유병률이 높은 질병에서부터 기존 일상에서 불편을 느끼는 가벼운 질환으로까지의 적용 범위가 확장되고 있다. 이에 따라서 의료기관들은 환자에 대한 라이프로그 데이터, 의료 이미지 데이터, 병원 진료기록 등을 수집하고 분석하고 있다[1, 2]. 특히 수집된 헬스케어 데이터는 텍스트 및 이미지, 영상과 같은 비정형 데이터로 구성되어 있다. 그러나 질병의 종류에 따라 양성 데이터와 음성 데이터는 비율의 불균형이 존재한다. 이는 헬스케어 데이터 분석을 통한 질병 예측 모델의 성능을 저하시키는 요인으로 작용한다.

신체의 이상이 없는 대부분의 여성은 사춘기에 초경을 시작하여 30-40년간의 장기간 동안 월경을 한다. 이 기간 동안 대부분의 여성은 최소 한 번 이상의 월경불순을 경험한다. 월경불순은 자궁 기능의 질환, 심리 및 정신적 원인, 신체 밸런스의 붕괴 등으로 다양한 원인을 가져 진단 및 예측에 어려움이 있다. 지속적인 월경불순은 배란 장애가 동반되기 때문에 여성호르몬 과잉 상태가 지속되어 여성암 및 자궁암의 원인이 된다. 본 연구에서는 질병관리청 국민건강영양조사에서 제공하는 원시자료에서 월경불순의 원인이 되는 특징을 분석하고 추출된 특징들을 통해 머신러닝을 활용한 생리불순 이진 분류 모델을 제안한다. 그러나 국민건강영양조사 원시자료에는 결측값과 데이터 불균형이 존재한다. 이는 데이터를 이용하여 학습하는 모델 성능의 신뢰도를 저하한다[3].

따라서 본 연구에서는 불균형 클래스에서 결측값 대치와 오버샘플링 알고리즘을 이용하여 AutoML 기반 이진 분류 모델의 성능 향상을 제안한다. MICE(Multiple Imputation by Chained Equations)를 활용한 데이터 결측값 대치를 이용하여 데이터 전처리를 진행한다[4]. 헬스 데이터의 경우 특정 질환에 대하여 양성 데이터와 음성 데이터의 불균형이 심하다. 이는 양성인 환자보다 음성인 환자의 데이터 비율이 부족하기 때문에 발생한다. 따라서 양성 데이터와 음성 데이터의 비율을 맞추어 데이터 불균형 문제를 해결하기 위해 오버 샘플링 알고리즘인 SMOTE(Synthetic Minority Over-sampling Technique)를 통해 양성 데이터와 음성 데이터의 비율을 맞춘다[5]. 데이터 전처리를 통해 가공된 데이터는 AutoML(Automated machine learning)을 사용하여 데이터에 적합한 모델 학습을 비교 분석한다[6]. 사용자는 자신의 헬스 데이터에 따라 분류 모델을 통해 월경불순이 예측 가능하다. 또한 월경불순이 지속됨에 따라 유발가능성 있는 여성암 및 자궁질환에 대한 예방책으로 사용한다[7].

2. 관련 연구

2.1 MICE 함수를 활용한 결측값 대치

MICE(Multiple Imputation by Chained Equations)는 연쇄방정식을 통해 자료의 결측치를 대체하는 접근법이다. Fig. 1은 MICE 함수를 활용한 결측값 대치 과정을 나타낸다.

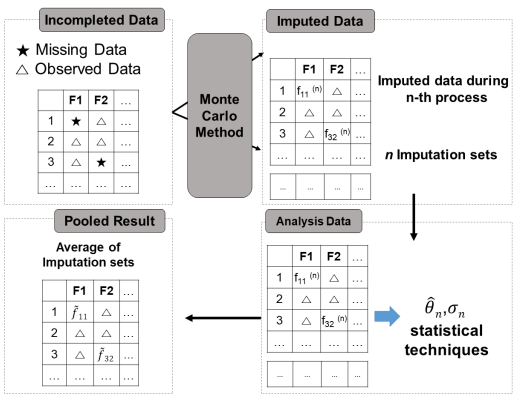

Fig. 1. Process of Multiple Imputation by Chained Equations

Fig. 1에서 MICE 함수는 결측치가 존재하는 데이터 셋으로 시작하여 완전한 데이터를 담고 있는 객체를 반환한다. 결측 데이터가 존재하는 데이터 셋에 대해 무작위로 추출된 난수를 이용하여 대체값을 계산하는 Monte Carlo를 사용한다[8]. 이를 통해 결측 데이터는 시뮬레이션을 통하여 5~10개의 완전한 데이터 셋을 얻는다. 각각의 데이터 셋에 대해 표준적 통계모델을 적용시켜 해당되는 모수 및 표준편차 등을 구한다. 통계적 데이터를 통해 각 모델에 대한 신뢰구간 및 결과값이 산출된다. 마지막으로 각 변수의 확률분포를 통합하여 목표값의 확률분포를 추정하는 Pool 함수를 사용한다. 따라서 최종 모델의 표준오차와 p 값은 소수데이터 다중 대치에 의해 생산된 불확실성을 반영한다. 이러한 알고리즘은 누락된 값을 하나의 통계적 데이터(중앙값, 평균값)으로 채워 넣는 단순대치법(Single Imputation)

과 비교하여 표준오차에 대한 과소평가 및 부정확한 p-value 값이 산출되는 것을 상대적으로 낮춘다[9].

2.2 SMOTE를 활용한 데이터 불균형 처리

SMOTE 알고리즘은 기존데이터를 단순히 복제하는 것이 아니라 소수 클래스의 데이터들을 서로 보완하여 새로운 데이터를 합성하는 샘플링 방식이다. 식(1)은 기존의 소수데이터를 보간하여 새로운 데이터를 생성하는 방법이다.

$$x_{diff} = x_i + (\hat{x_i} - x_i) \times rand[0,1] \quad (1)$$

독립변수 x_i를 갖는 클래스 I인 데이터를 하나 선정한다. 선정된 데이터로부터 얻을 수 있는 k개의 클래스 I인 근접한 이웃 데이터를 탐색한다. 탐색된 데이터의 독립변수를 $\hat{x_i}$라고 한다. 새롭게 생성될 점 x_{diff}는 x_i와 $\hat{x_i}$를 잇는 선분에서 임의로 뽑은 임의의 한 점이 된다. $rand[0,1]$은 0과 1 사이의 무작위 수이다. x_{diff}에서 동일 클래스의 데이터가 추가되고, 사용자의 지정값 만큼의 반복을 통하여 데이터의 비율을 맞춘다[10]. 따라서 소수 데이터가 다수데이터와 동등하게 샘플링 되는 것이 가능하다. Fig. 2는 SMOTE 알고리즘의 작동방식이다.

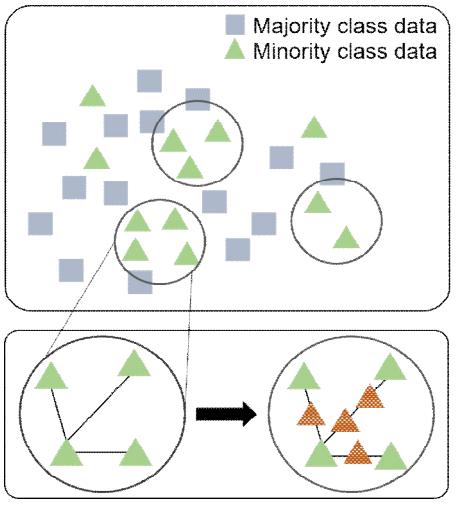

Fig. 2. Operation of SMOTE Algorithm

먼저 소수 데이터에 대해 k개의 가장 가까운 이웃을 선택한다. 첫 번째 선택된 데이터를 기준으로 인접 데이터 사이의 중간에 선형 보간법을 사용하여 데이터를 생성한다. 업 샘플링과 같이 소수 데이터를 반복 및 복제하여 적정 비율로 설정하는 것이 아닌 클래스의 이동에서 기인하는 변화를 통해 데이터를 합성하기 때문에 과적합 문제를 해결한다.

2.3 AutoML을 활용한 모델 생성

AutoML은 시간 소모적이고 반복적인 기계학습 모델의 개발 작업을 자동화하는 프로세스이다. Fig. 3은 전처리된 데이터를 사용하여 생성된 이진분류기를 나타낸다. 입력값과 파라미터 값의 조절을 통해 여러 모델을 데이터의 형태에 맞게 자동으로 생성한다. 입력된 데이터들이 회귀 및 분류 모델로 평가되며 가장 높은 성능을 가진 모델들이 차례로 산출된다. 산출된 모델들은 기준값에 따라 조합되어 새로운 앙상블 모델을 생성한다.

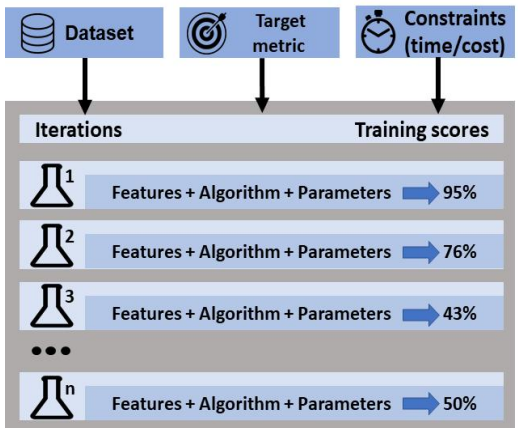

Fig. 3. Binary Classifier Generated using Preprocessed Data

3. 불균형 클래스에서 AutoML 기반 분류 모델의 성능 향상을 위한 데이터 처리

3.1 데이터 수집 및 전처리

질병관리본부가 제공하는 국민건강영양조사 원시자료에서 월경불순인 여성과 아닌 여성에 대한 데이터를 수집한다[11]. 월경불순의 경우 정신적 스트레스, 과도한 비만, 골밀도 수치 저하 등의 다양한 원인으로 발생한다[12,13]. 본 연구에서는 T-score, BMI(Body Mass Index), 개인의 정신적 스트레스 정도를 특성으로 둔다. T-score는 대퇴골, 요추, 대퇴골 경부에 대한

최대 골밀도 연령군의 표준편차이다. 정신적 스트레스의 강도는 범주형 데이터이다. 모델 학습 알고리즘이 범주형 데이터를 인식할 수 없기 때문에 모든 범주형 변수를 숫자형으로 인코딩하는 전처리 작업이 필요하다. 따라서 범주형 데이터인 스트레스 정도는 Onehotencoding을 통해 전처리한다. 골격도와 BMI는 평균을 0, 분산을 1로 변경하는 Standscaler 함수를 사용하여 같은 범위를 갖게 한다. 이는 특성 값의 범위가 달라 모델의 알고리즘이 데이터를 학습하는 과정에서 0으로 수렴하거나 무한으로 발산하는 경우를 방지하기 위함이다[14].

3.2 MICE 결측치 대치와 오버샘플링을 활용한 데이터 전처리

입력 데이터는 2703개의 행으로 구성되어 있다. 이 중 결측치가 포함된 행은 612개로 전체 데이터의 22%를 차지한다. 입력 데이터 중 골격도 데이터는 결측치가 존재하기 때문에 MICE를 활용하여 결측치를 대치하여 정제된 데이터 셋을 구성한다. 결측치가 처리된 데이터 셋은 생리불순이 아닌 여성과 생리불순인 여성의 데이터비율에 불균형이 존재한다. Fig. 4는 생리불순 데이터의 전처리 과정을 나타낸다. 생리불순 양성 데이터는 381개, 음성 데이터는 2322개로 구성된다. 데이터 불균형을 개선하기 위한 오버샘플링 알고리즘인 SMOTE를 사용하여 비율을 조정한다. SMOTE 알고리즘이 적용된 데이터의 비율은 양성 데이터 1000개, 음성데이터 2322개로 약 1:2로 조정된다.

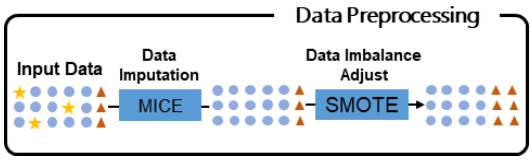

Fig. 4. Preprocessing of Dysmennorrhea Data

3.3 AutoML기반 분류 모델 생성

Fig. 5는 AutoML을 기반으로 MICE 결측값 대치와 SMOTE 알고리즘을 사용하여 데이터 불균형이 해소된 데이터를 이용하여 모델을 생성하는 과정을 나타낸다. 전처리 된 데이터는 총 20개의 회귀 및 분류 모델에 적용된다. 모델들은 생리불순인 여성과 생리불순이 아닌 여성들의 특성값을 학습한다. 학습이 완료된 모델들은 각 성능이 산출되며 사용자가 의도하는 성능 평가 지표를 기준으로 순차적으로 정렬된다. 모델들 중 가장 높은 성능을 기록한 모델 3개를 조합하여 앙상블 모델을 생성한다. 이는 여러 모델을 직접 실험해보는 시간 및 리소스를 절약하고 각 모델에 대한 하이퍼 파라미터를 수정하지 않는 장점을 가진다.

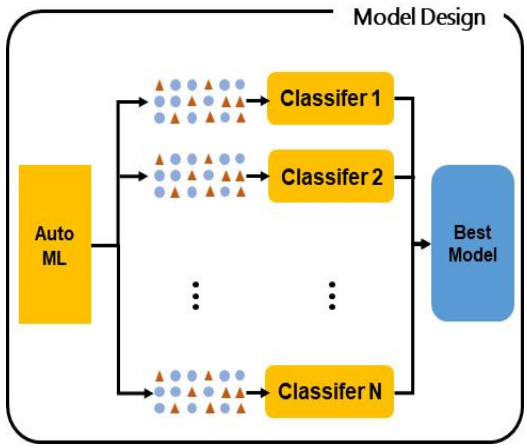

Fig. 5. Process of Creating a Model using Preprocessed Data

4. 결과 및 성능 평가

제안된 AutoML을 활용한 생리불순 이진분류 모델은 두 가지 성능평가를 진행한다. 첫째로 산출된 모델은 정확도에 의해 평가된다. 정확도는 직관적으로 산출된 모델의 성능을 평가할 때 사용된다. 식(2)는 정확도의 산출식을 나타낸다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

식(2)에서 TP는 True Positive, TN은 True Negative, FP는 False Positive, FN은 False Negative를 의미한다. 그러나 정확도는 입력 데이터 도메인이 불균형한 모델의 경우 편중이 발생하는 한계를 가진다. 이는 예측 모델의 신뢰도를 감소시킨다. 이를 보완하기 위해 F1-score를 사용한다.

$$F1 - score = 2 \times \frac{recall * precision}{recall + precision} \quad (3)$$

$$\mathrm{Precision} = \frac{TP}{TP+FP}\ ,\ Recall = \frac{TP}{TP+FN} \quad (4)$$

두 번째 성능평가는 F1-score를 통해 진행한다. 식(3)은 F1-score의 산출식을 나타낸다. F1-score는 정밀도와 재현율의 조화평균을 이용한 성능지표다. 즉, 두 지표를 모두 균형있게 반영하여 모델의 편향을 확인한다. 따라서 F1-score를 높이기 위해서는 정밀도와 재현율의 균일한 값이 요구된다. 정밀도와 재현율은 식(4)으로 표현된다. 정밀도는 모델이 판단하는 생리불순인 여성 중 실제 양성인 비율을 의미한다. 재현율은 실제로 여성이 생리불순인 경우의 수 중에서 실제 양성인 비율을 의미한다[15]. 일반적으로 정밀도와 재현율은 Trade-off의 관계를 가진다. Table 1은 전처리가 완료된 데이터를 사용하여 AutoML로 생성한 모델들의 F1-score를 기준으로 기록된 성능평가 결과이다.

Table 1. Performance Evaluation Result of Single Classifier Generated by AutoML

Index	Recall	Prec.	F1-score
Extra Trees	0.9246	0.8282	0.8737
Random Forest	0.9138	0.8413	0.8760
Light Gradient	0.9181	0.8537	0.8747
Decision Tree	0.9246	0.8266	0.8728
Gradient Boosting	0.9246	0.8346	0.8773
...	...	...	...
Mean	0.9211	0.8369	0.8769

Table 2. Performance Result for Ensemble Model by AutoML

Accuracy	Recall	Precision	F1-score
0.8758	0.9175	0.8468	0.8807

기록된 총 20개의 모델 중 F1-score가 높게 평가된 3개의 모델을 앙상블하여 최종모델을 산출한다. Table 2는 앙상블모델의 성능 평가표이다. 산출된 상위 3개의 모델인 Extra Trees, Random Forest, Light Gradient의 F1-score가 각각 87.37%, 87.60%, 87.47% 이다. 이에 비하여 앙상블 모델의 F1-score는 88.07%로 기존 분류 모델들과 비교하여 우수한 성능을 보인다. Table 3은 MICE와 SMOTE를 적용한 데이터와 Raw Data의 F1-score를 Extra Trees Classification과 앙상블모델에 대해 비교한 표이다. 두 모델 각각 전처리 후 Raw Data에 비해 F1-score가 65.9%p, 71.9%p의 성능향상을 보인다. 전처리 된 데이터를 사용한 AutoML 기반 앙상블 모델은 Extra Trees Classification에 비해 약 7%p의 성능향상을 보인다. 이는 MICE 결측값 대치와 SMOTE 오버샘플링 알고리즘을 활용한 전처리 방식이 모델성능을 향상시켰음을 보인다.

Table 3. Overall Performance Comparison between the Proposed Method and the Existing Method

	Raw Data	Using MICE and SMOTE
Extra Trees	0.1476	0.8066
Ensemble Model	0.1610	0.8807

5. 결론

본 연구에서는 불균형 클래스에서 AutoML 기반 분류 모델의 성능 향상을 위한 데이터 처리를 제안하였다. 결측치 대치와 오버 샘플링을 통해 정제된 데이터를 사용하여 AutoML 기반의 생리불순 분류 모델을 개발하였다. 기존의 원시자료는 환자의 불응답 및 기록되지 않은 값인 결측값이 존재한다. 이를 MICE를 활용하여 결측값을 대치한다. 또한 소수 데이터가 존재하여 신뢰있는 모델 성능을 산출하는데 한계가 존재한다. 따라서 데이터 불균형을 해소하는 오버 샘플링을 사용하여 소수데이터를 가공한다. 가공된 데이터가 기존데이터에 비하여 상향된 모델 성능을 산출하는데 기여한다. 기존 데이터의 경우 F1-score 약 14%로 신뢰가능하지 않은 값을 산출한다. 제안한 방법의 성능 평가를 진행한 결과 정확도는 약 87.5%이고 F1-score는 약 88%로 신뢰가능한 성능을 보였다. 이는 샘플의 불균형이 존재하는 헬스케어 데이터에서 MICE 결측치 처리와 SMOTE 오버샘플링을 통해 성능 지표를 상향하는 것이 가능함을 보인다.

REFERENCES

[1] J. C. Kim & K. Chung. (2020). Hybrid Multi-Modal Deep Learning using Collaborative Concat Layer in Health Bigdata. *IEEE Access, 8,* 192469-192480. DOI : 10.1109/ACCESS.2020.3031762

[2] H. Asri, H. Mousannif, H. Al Moatassime & T. Noel. (2015, June). Big data in healthcare: challenges and opportunities. *In International*

Conference on Cloud Technologies and Applications (CloudTech), 1-7.

[3] J. C. Kim & K. Chung. (2020). Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. *IEEE Access, 8,* 104933-104943.
DOI : 10.1109/ACCESS.2020.2997255

[4] S. V. Buuren & K. Groothuis-Oudshoorn. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software,* 1-68.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16,* 321-357.
DOI : 10.1613/jair.953

[6] A. Truong, A. Walters, J. Goodsitt, K. Hines, C. B. Bruss & R. Farivar. (2019). Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. *In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence(ICTAI),* 1471-1479.
DOI : 10.1109/ICTAI.2019.00209

[7] D. J. Lee, J. S. Kang, M. J. Kim, J. W. Baek & K. Chung. (2021). Data Imbalance Processing through Over-sampling in Binary Classification Model. *Korean Society For Internet Information Spring Conference,* 77-78.

[8] Y. A. Shreider (2014). *The Monte Carlo method: the method of statistical trials*(Vol. 87). Elsevier.

[9] I. R. White, P. Royston & A. M. Wood. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine, 30(4),* 377-399.
DOI : 10.1002/sim.4067

[10] C. Gong & L. Gu. (2016). A novel SMOTE-based classification approach to online data imbalance problem. *Mathematical Problems in Engineering,* 1-14.

[11] The Fifth Korea National Health and Nutrition Examination Survey (KNHANES V-2). (2015). Korea Centers for Disease Control and Prevention.

[12] B. L. Drinkwater, B. Bruemner & C. H. Chesnut. (1990). Menstrual history as a determinant of current bone density in young athletes. *Jama, 263(4),* 545-548.
DOI : 10.1001/jama.1990.03440040084033

[13] Y. Liu, E. B. Gold, B. L. Lasley & W. O. Johnson. (2004). Factors affecting menstrual cycle characteristics. *American journal of epidemiology, 160(2),* 131-140.
DOI : 10.1093/aje/kwh188

[14] J. Hao & T. K. Ho. (2019). Machine learning made easy: A review of scikit-learn package in Python programming language. *Journal of Educational and Behavioral Statistics, 44(3),* 348-361.
DOI : 10.3102/1076998619832248

[15] S. E. Ryu, D. H. Shin & K. Chung. (2020). Prediction model of dementia risk based on XGBoost using derived variable extraction and hyper parameter optimization. *IEEE Access, 8,* 177708-177720.
DOI : 10.1109/ACCESS.2020.3025553

이 동 준(Dong-Joon Lee) [학생회원]

·2016년 3월 ~ 현재 : 경기대학교 글로벌어문학과, AI컴퓨터공학부(복수전공) (학부생)
·2021년 1월 ~ 현재 : 경기대학교 컴퓨터과학과 데이터마이닝 연구실 연구원

·관심분야 : 데이터 마이닝, 데이터 분석, 지능시스템, 헬스케어
·E-Mail : dongzza97@kyonggi.ac.kr

강 지 수 (Ji-Soo Kang) [학생회원]

·2020년 2월 : 경기대학교 컴퓨터공학부 (공학사)
·2020년 3월 ~ 현재: 경기대학교 컴퓨터과학과 (석사과정)
·2017년 12월 ~ 현재: 경기대학교 컴퓨터과학과 데이터마이닝 연구실 연구원

·관심분야 : 데이터 마이닝, 데이터 분석, 지능시스템, 헬스케어, 의료 딥 러닝
·E-Mail : dm.jskang@kyonggi.ac.kr

정 경 용(Kyungyong Chung) [정회원]

·2000년 2월 : 인하대학교 전자계산공학과 (공학사)
·2002년 2월 : 인하대학교 전자계산공학과 (공학석사)
·2005년 8월 : 인하대학교 컴퓨터정보공학부 (공학박사)

·2006년 3월 ~ 2017년 2월 : 상지대학교 컴퓨터정보공학부 교수
·2017년 3월 ~ 현재 : 경기대학교 컴퓨터공학부 교수
·관심분야 : 데이터마이닝, 헬스케어, 빅데이터, 지능시스템, 인공지능, HCI, 정보검색, 추천 시스템
·E-Mail : dragonhci@hanmail.net