2024 한국정보기술학회 하계종합학술대회 및 대학생논문경진대회

불균형 데이터를 위한 샘플링 기반 심장질환 진단 모델

Sampling-Based Cardiac Diagnosis Model for Imbalanced Data

박지호·심수지·채유진·신제용

한국외국어대학교
조선대학교
인하대학교
독립 연구가

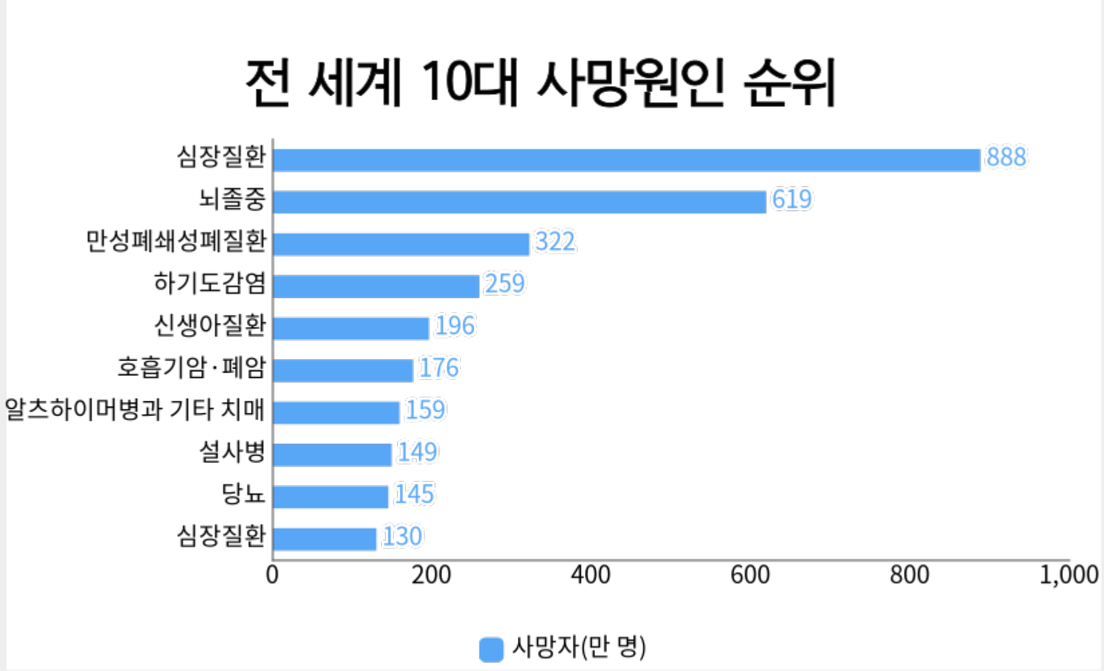
Contents

1. 연구 배경

1. 연구 배경

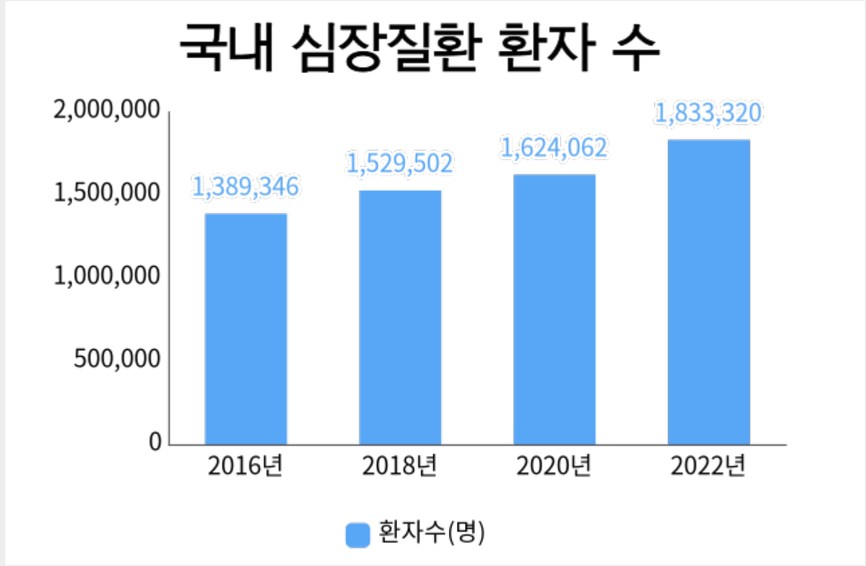
세계·국내 심장질환의 높은 유병률과 이로 인한 사회경제적 비용 상승

- 2019년 기준, 심장질환은 전 세계 사망원인 1위 질환



출처: The top 10 causes of death (WHO, 2020)

- 2022년 심장질환 환자 수는 183만 3,320명으로 2018년 대비 **19.9% 증가**
- 진료비는 2조 5,391억원으로 2018년 대비 **38.5% 증가**



출처: 2022년 사망원인통계 결과 (통계청, 2023)

1. 연구 배경

전통적인 심장질환 진단 방법의 한계

1. 진단 정확성
- 심장질환 진단의 비특이성
  - 심장질환은 증상이 비특이적이며, 뚜렷한 증상 없이 진행되는 경우가 많다. 이는 정확한 진단과 조기진단을 어렵게 만든다.

2. 기술 접근성
- 비용 문제 및 전문가 의존성
  - 심장 초음파와 같은 고가의 진단 장비를 사용하고 결과를 해석하기 위해서는 전문 인력이 필요하며, 재정적, 기술적 자원이 부족한 환경에서는 이에 대한 접근성이 제한된다.

머신러닝을 통한 개선

1. 심장질환 증상 진단 정확도 향상
- 머신러닝을 통해 대규모 데이터셋을 분석하여 다양한 변수를 고려한 심장질환 징후를 파악하고, 질병의 조기진단 가능성을 높임

2. 진단 기술에 대한 접근성 향상
- 원격 데이터 수집 및 분석이 가능하여 의료 접근성의 격차를 줄이고, 고가의 의료 장비에 대한 의존도를 낮출 수 있다.

1. 연구 배경

세계·국내 심혈관 질환의 높은 유병률과 이로 인한 사회경제적 비용 상승

전통적인 심장질환 진단 방법의 한계

- **머신러닝**을 활용하여 심장질환 진단 및 조기진단의 확률 높임
- 의료 데이터에서 발생하는 클래스 불균형 문제를 **리샘플링 기법**을 적용하여 개선

2. 데이터 수집 및 전처리

2. 데이터 수집 및 전처리

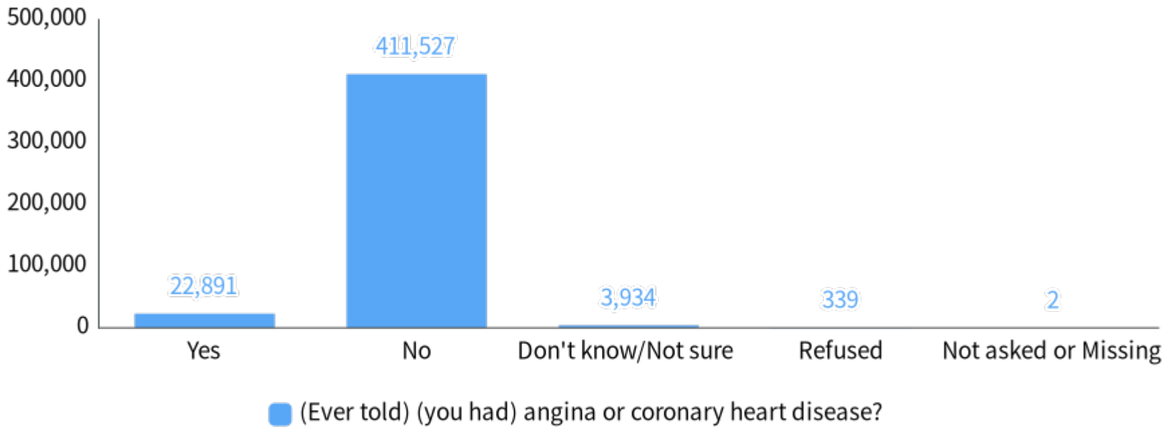
1) 데이터셋

Behavorial Risk Factor Surveillance System (BRFSS) 2021 dataset

- 438,693 x 303
- 미국 질병통제예방센터(CDC)가 주관하는 연례 설문조사
- 미국 전역의 성인들을 대상으로 실시
- 개인의 건강 및 생활습관에 대한 정보를 전화 설문을 통해 수집
- 공중보건 및 연구에서 매우 중요한 역할

Label: Ever Diagnosed with Angina or Coronary Heart Disease
Section Name: Chronic Health Conditions
Core Section Number: 7
Question Number: 2
Column: 120
Type of Variable: Num
SAS Variable Name: CVDCRHD4
Question Prologue:
Question: (Ever told) (you had) angina or coronary heart disease?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	22,891	5.22	3.83
2	No	411,527	93.81	95.39
7	Don't know/Not sure	3,934	0.90	0.72
9	Refused	339	0.08	0.07
BLANK	Not asked or Missing	2	.	.



2. 데이터 수집 및 전처리

2) 전처리

① 종속 변수인 심장질환 유무를 제외한 나머지 302개의 독립 변수 중, 심장질환과 연관성이 높은 21개의 변수 선정

- 프레밍험(Framingham) 위험지수: 나이, 성별, 콜레스테롤, 혈압, 흡연
- 선행 연구
- 변수 선정: 고혈압 여부, 고콜레스테롤 여부, 최근 콜레스테롤 수치 검사 여부, 체질량지수(BMI), 흡연 여부, 뇌졸중 여부, 당뇨병 여부, 신체활동 수준, 과일 섭취 수준, 채소 섭취 수준, 과도한 음주 여부, 의료보험 가입 여부, 의료비 부담 여부, 전반적인 건강 상태, 정신건강 상태, 신체건강 상태, 걷기 어려움 여부, 성별, 연령대, 교육 수준, 가구 소득 수준

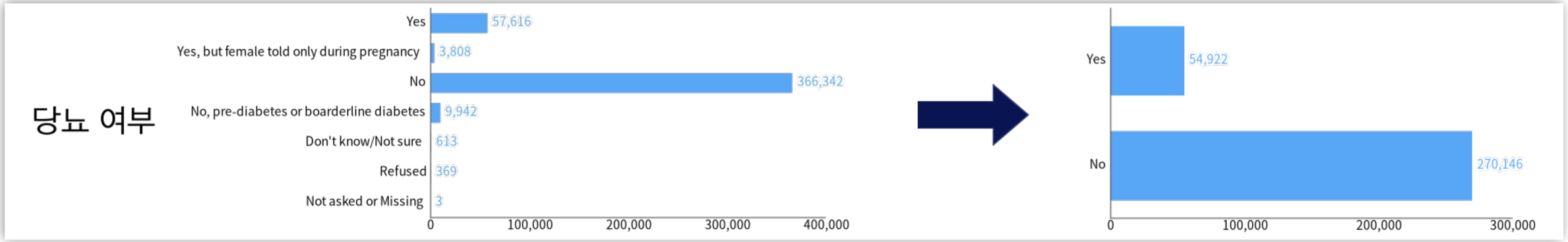
변수명	의미
_RFHYPE6	고혈압 여부
_TOLDHI3	고콜레스테롤여부
_CHOLCHK3	최근 콜레스테롤수치 검사 여부
_BMI5	체질량지수(BMI)
SMOKE100	흡연 여부
CVDSTRK3	뇌졸중 여부
DIABETE4	당뇨병 여부
_TOTINDA	신체활동 수준
_FRTLT1A	과일 섭취 수준
_VEGLT1A	채소 섭취 수준
_RFDRHV7	과도한 음주 여부
_HLTHPLN	의료보험 가입 여부
MEDCOST1	의료비 부담 여부
GENHLTH	전반적인 건강 상태
MENTHLTH	정신건강 상태
PHYSHLTH	신체건강 상태
DIFFWALK	걷기 어려움 여부
_SEX	성별
_AGEG5YR	연령대
EDUCA	교육 수준
INCOME3	가구 소득 수준

2. 데이터 수집 및 전처리

2) 전처리

② 범주형 데이터

- 불명확한 답변, 결측치 제거
- 모름, 답변 거부 제거



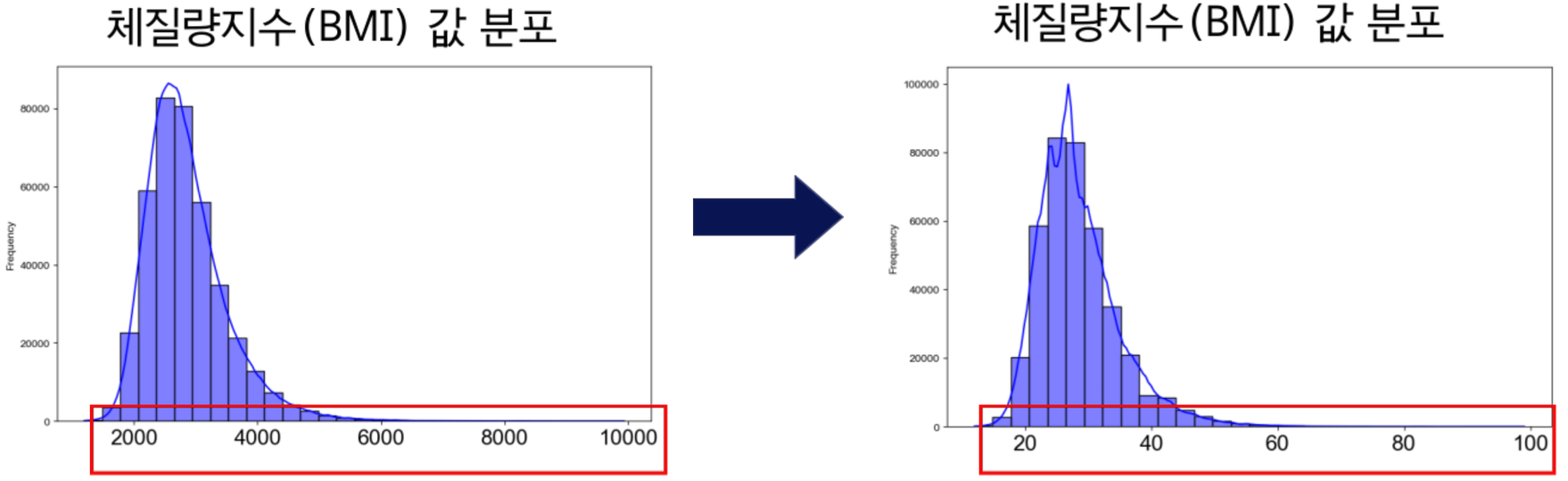
2. 데이터 수집 및 전처리

2) 전처리

③ 수치형 데이터 (BMI)

- 네 자리 숫자로 기록된 BMI 값에 대하여 100으로 나눈 후, 정수형으로 변환

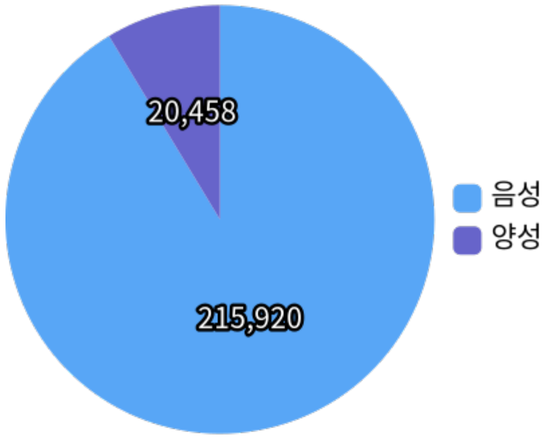
체질량지수
(BMI)



3. 모델 설계

3. 모델 설계

1) Resampling



- 음성 데이터가 양성 데이터의 약 10배인 **매우 불균형**한 데이터셋
- 클래스 불균형에 따른 모델 성능 저하 방지를 위해 **Resampling 기법** 사용

- Undersampling
  - Random Undersampling (RUS)
  - Tomek Links
- Oversampling
  - SMOTE
  - Borderline-SMOTE
  - ADASYN
- Undersampling & Oversampling
  - SMOTETOMEK
  - SMOTEENN

3. 모델 설계

2) Algorithm

Logistic Regression
- 결과 변수가 이진 분류일 경우에 많이 사용되는 선형 분류 알고리즘

Decision Tree
- 의사 결정 규칙을 나무 구조로 도표화하여 분류와 예측 수행

XGBoost
- 여러 개의 Decision Tree 모델을 Boosting 앙상블로 구현

LightGBM
- 여러 개의 Decision Tree 모델을 Boosting 앙상블로 구현
- GBDT 알고리즘의 단점을 보완하여 학습시간 단축

4. 실험 결과

4. 실험 결과

1) Recall 중점으로 모델 평가
2) Recall이 0.8 이상이면서 Precision이 0.2 이상인 모델

- Precision = $\frac{TP}{TP + FP}$
- Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$
- AUC
- Recall = $\frac{TP}{TP + FN}$
- F1-Score = $\frac{2 \times Precision \times Recall}{Precision + Recall}$

4. 실험 결과

1) Recall 중점으로 모델 평가
2) Recall이 0.8 이상이면서 Precision이 0.2 이상인 모델

Recall 이 0.95 이상인 모델의 평가지표

Resample	Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
RUS	LightGBM	0.4046	0.1248	0.9778	0.2213	0.6641
RUS	XGBoost	0.4377	0.1298	0.9643	0.2289	0.6760
Average		**0.7394**	**0.2137**	**0.6927**	**0.3125**	**0.7184**

4. 실험 결과

1) Recall 중점으로 모델 평가
2) Recall이 0.8 이상이면서 Precision이 0.2 이상인 모델

Recall 0.8 이상, Precision 0.2이상인 모델의 평가지표

Resample	Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Tomek Links	LightGBM	0.7401	0.2228	0.8051	0.3490	0.7695
SVMSMOTE	LightGBM	0.7344	0.2195	0.8098	0.3454	0.7685
SVMSMOTE	XGBoost	0.7374	0.2209	0.8051	0.3467	0.7681
SMOTEENN	Logistic Regression	0.7019	0.2027	0.8336	0.3261	0.7615
SMOTETOMEK	LightGBM	0.7220	0.2136	0.8252	0.3394	0.7687
SMOTE	LightGBM	0.7201	0.2126	0.8265	0.3382	0.7682
BorderlineSMOTE	LightGBM	0.7226	0.2141	0.8256	0.3400	0.7692
ADASYN	LightGBM	0.7199	0.2126	0.8271	0.3382	0.7684

5. 결론 및 향후 연구

5. 결론 및 향후 연구

- 심장질환의 진단 정확성을 향상시키기 위해 머신러닝 기법 활용
- Resampling 기법 적용하여 클래스 불균형 문제에 대응
- **SMOTEENN 기법을 적용한 Logistic Regression 모델**이 Recall을 중심으로 전반적인 성능 지표에서 우수한 결과

- **Tomek Links 방법을 적용한 LightGBM 모델**이 Recall을 제외한 평가지표에서 우위
- Gridsearch 등의 하이퍼 파라미터 최적화 기법을 통해 성능 향상 기대
- Resampling 기법과 알고리즘 간의 상호작용 연구

감사합니다