



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0130992
(43) 공개일자 2024년08월30일

(51) 국제특허분류(Int. Cl.)
G16H 50/20 (2018.01) G06N 20/20 (2019.01)
G16H 50/50 (2018.01)
(52) CPC특허분류
G16H 50/20 (2018.01)
G06N 20/20 (2021.08)
(21) 출원번호 10-2023-0024136
(22) 출원일자 2023년02월23일
심사청구일자 2023년02월23일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대 학교)
(72) 발명자
곽경섭
인천광역시 연수구 해돋이로84번길 10, 604동 80 2호(송도동, 송도풍림아이원6단지아파트)
(74) 대리인
양성보

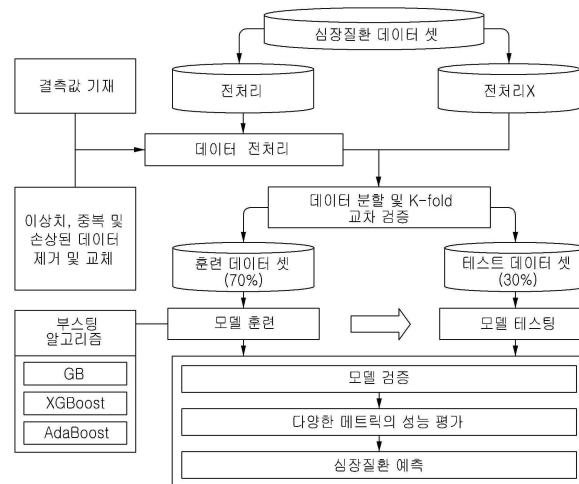
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 부스팅 알고리즘을 활용한 심장질환 예측을 위한 새로운 앙상블 학습 방법 및 시스템

(57) 요약

부스팅 알고리즘을 활용한 심장질환 예측을 위한 새로운 앙상블 학습 방법 및 시스템이 개시된다. 일 실시예에 따른 심장질환 예측 시스템에 의해 수행되는 심장질환 예측 방법은, 심장질환을 예측하기 위해 구성된 앙상블 모델에 심장질환 데이터를 입력받는 단계; 및 상기 앙상블 모델을 이용하여 상기 입력받은 심장질환 데이터로부터 심장질환을 예측하는 단계를 포함하고, 상기 앙상블 모델은, 부스팅 알고리즘이 적용되어 심장질환 데이터 셋에 포함된 속성 정보를 이용하여 환자별 특징 정보가 학습됨에 따라 심장질환을 예측하도록 구축된 것일 수 있다.

대표도



(52) CPC특허분류

G16H 50/50 (2018.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711166941
과제번호	2020R1A2B5B02002478
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	중견후속연구
연구과제명	[Ezbaro] 빅데이터 분석 기반 퍼지 온톨로지 및 기계학습을 활용한 IoT u-헬스케어
시스템 핵심기술 개발 및 설계	
기 여 율	1/1
과제수행기관명	인하대학교
연구기간	2022.03.01 ~ 2023.02.28

명세서

청구범위

청구항 1

심장질환 예측 시스템에 의해 수행되는 심장질환 예측 방법에 있어서,
심장질환을 예측하기 위해 구성된 앙상블 모델에 심장질환 데이터를 입력받는 단계; 및
상기 앙상블 모델을 이용하여 상기 입력받은 심장질환 데이터로부터 심장질환을 예측하는 단계
를 포함하고,
상기 앙상블 모델은, 부스팅 알고리즘이 적용되어 심장질환 데이터 셋에 포함된 속성 정보를 이용하여 환자별
특징 정보가 학습됨에 따라 심장질환을 예측하도록 구축된 것을 특징으로 하는 심장질환 예측 방법.

청구항 2

제1항에 있어서,
상기 앙상블 모델은,
그레디언트(gradient) 부스트, AdaBoost 또는 XGBoost를 포함하는 둘 이상의 부스팅 알고리즘이 적용된 것을 특
징으로 하는 심장질환 예측 방법.

청구항 3

제1항에 있어서,
심장질환을 예측하기 위한 데이터 셋을 이용하여 심장질환 예측을 위해 구성된 앙상블 모델을 학습시키는 단계
를 더 포함하는 심장질환 예측 방법.

청구항 4

제3항에 있어서,
상기 학습시키는 단계는,
심장질환을 예측하기 위한 데이터 셋에 대해 전처리 과정을 수행 또는 전처리 과정을 미수행하고, 상기 전처리
과정이 수행된 데이터 셋 또는 상기 전처리 과정이 미수행된 데이터 셋 중 하나 이상의 데이터 셋을 획득하여
데이터 분할 및 K-fold 교차 검증을 수행하고, 상기 수행된 데이터 분할 및 K-fold 교차 검증을 통해 획득된 최
종 데이터 셋에 대해 기 설정된 비율로 훈련 데이터 셋과 테스트 데이터 셋을 분할하는 단계
를 포함하는 심장질환 예측 방법.

청구항 5

제4항에 있어서,
상기 학습시키는 단계는,
상기 분할된 훈련 데이터 셋을 이용하여 각각의 부스팅 알고리즘이 적용된 모델을 훈련시키고, 상기 훈련된 모
델에 상기 분할된 테스트 데이터 셋을 이용하여 상기 훈련된 모델을 테스트한 후, 모델 검증 및 모델 성능을 평
가하여 심장질환 예측을 위한 앙상블 모델을 도출하는 단계
를 포함하는 심장질환 예측 방법.

청구항 6

제4항에 있어서,

상기 전처리 과정은,

상기 심장질환 데이터 셋에 대한 전처리를 통해 결측값을 식별함에 따라 결측값 보정 방법을 이용하여 상기 식별된 결측값을 다른 값으로 변경하고,

상기 심장질환 데이터 셋에 있는 이상치를 검출함에 따라 사분위간 범위 방법을 이용하여 상기 검출된 이상치를 다른 값으로 변경하는 것을 특징으로 하는 심장질환 예측 방법.

청구항 7

심장질환 예측 시스템에 있어서,

심장질환을 예측하기 위해 구성된 앙상블 모델에 심장질환 데이터를 입력받는 데이터 입력부; 및

상기 앙상블 모델을 이용하여 상기 입력받은 심장질환 데이터로부터 심장질환을 예측하는 심장질환 예측부를 포함하고,

상기 앙상블 모델은, 부스팅 알고리즘이 적용되어 심장질환 데이터 셋에 포함된 속성 정보를 이용하여 환자별 특징 정보가 학습됨에 따라 심장질환을 예측하도록 구축된 것을 특징으로 하는 심장질환 예측 시스템.

발명의 설명

기술 분야

[0001]

아래의 설명은 심장질환 예측 기술에 관한 것이다.

배경 기술

[0003]

심장병은 전 세계적으로 인간의 삶에 영향을 미치는 위험 중 하나로 간주되고 있다. 국제 의료 기관의 통계 보고서에 따르면, 심혈관 질환으로 인해 2019 년 1,790 만명(모든 전 세계 사망자의 32%)이 사망하였으며, 2030 년까지 2,300만명으로 증가할 것으로 추정된다. 모든 심혈관 질환 사망 중 85%는 심장병과 뇌졸중으로 인한 것이었다. 미래에 조기 사망을 줄이기 위해서는 초기 단계에서 심장병을 감지하고 예측하는 것이 필요하다. 심장 관련 질병의 위험과 진행은 나이, 생활 양식의 변화, 식량 습관 및 건강 관리 센터 입원과 같은 빠르게 성장하는 사회 경제적 원인과 같은 요인에 따라 달라진다. 또한, 심장 관련 문제로 인한 다른 위험 요인으로는 고혈압, 포도당 수준 상승, 고급 혈액 지질, 비만 및 과체중이 있다. 이에, 심장 관련 질병을 보다 잘 예측하여 사전에 예방하고 예방 조치를 취할 수 있도록 컴퓨터 지능 기술을 탐색하는 것이 필요하다.

발명의 내용

해결하려는 과제

[0005]

부스팅 알고리즘이 적용된 앙상블 모델을 이용하여 심장질환을 예측하는 방법 및 시스템을 제공할 수 있다.

과제의 해결 수단

[0007]

심장질환 예측 시스템에 의해 수행되는 심장질환 예측 방법은, 심장질환을 예측하기 위해 구성된 앙상블 모델에 심장질환 데이터를 입력받는 단계; 및 상기 앙상블 모델을 이용하여 상기 입력받은 심장질환 데이터로부터 심장질환을 예측하는 단계를 포함하고, 상기 앙상블 모델은, 부스팅 알고리즘이 적용되어 심장질환 데이터 셋에 포함된 속성 정보를 이용하여 환자별 특징 정보가 학습됨에 따라 심장질환을 예측하도록 구축된 것일 수 있다.

[0008]

상기 앙상블 모델은, 그라디언트(gradient) 부스트, AdaBoost 또는 XGBoost를 포함하는 둘 이상의 부스팅 알고리즘이 적용된 것일 수 있다.

- [0009] 상기 심장질환 예측 방법은, 심장질환을 예측하기 위한 데이터 셋을 이용하여 심장질환 예측을 위해 구성된 앙상블 모델을 학습시키는 단계를 더 포함할 수 있다.
- [0010] 상기 학습시키는 단계는, 심장질환을 예측하기 위한 데이터 셋에 대해 전처리 과정을 수행 또는 전처리 과정을 미수행하고, 상기 전처리 과정이 수행된 데이터 셋 또는 상기 전처리 과정이 미수행된 데이터 셋 중 하나 이상의 데이터 셋을 획득하여 데이터 분할 및 K-fold 교차 검증을 수행하고, 상기 수행된 데이터 분할 및 K-fold 교차 검증을 통해 획득된 최종 데이터 셋에 대해 기 설정된 비율로 훈련 데이터 셋과 테스트 데이터 셋을 분할하는 단계를 포함할 수 있다.
- [0011] 상기 학습시키는 단계는, 상기 분할된 훈련 데이터 셋을 이용하여 각각의 부스팅 알고리즘이 적용된 모델을 훈련시키고, 상기 훈련된 모델에 상기 분할된 테스트 데이터 셋을 이용하여 상기 훈련된 모델을 테스트한 후, 모델 검증 및 모델 성능을 평가하여 심장질환 예측을 위한 앙상블 모델을 도출하는 단계를 포함할 수 있다.
- [0012] 상기 전처리 과정은, 상기 심장질환 데이터 셋에 대한 전처리를 통해 결측값을 식별함에 따라 결측값 보정 방법을 이용하여 상기 식별된 결측값을 다른 값으로 변경하고, 상기 심장질환 데이터 셋에 있는 이상치를 검출함에 따라 사분위간 범위 방법을 이용하여 상기 검출된 이상치를 다른 값으로 변경하는 것일 수 있다.
- [0013] 심장질환 예측 시스템은, 심장질환을 예측하기 위해 구성된 앙상블 모델에 심장질환 데이터를 입력받는 데이터 입력부; 및 상기 앙상블 모델을 이용하여 상기 입력받은 심장질환 데이터로부터 심장질환을 예측하는 심장질환 예측부를 포함하고, 상기 앙상블 모델은, 부스팅 알고리즘이 적용되어 심장질환 데이터 셋에 포함된 속성 정보를 이용하여 환자별 특징 정보가 학습됨에 따라 심장질환을 예측하도록 구축된 것일 수 있다.

발명의 효과

- [0015] 부스팅 알고리즘이 적용된 앙상블 모델을 통해 심장질환 예측의 정확도를 향상시킬 수 있다.

도면의 간단한 설명

- [0017] 도 1은 일 실시예에 있어서, 심장질환 예측 시스템의 구성을 설명하기 위한 블록도이다.
- 도 2는 일 실시예에 있어서, 심장질환 예측 방법을 설명하기 위한 흐름도이다.
- 도 3은 일 실시예에 있어서, 심장질환을 예측하기 위해 앙상블 모델을 학습시키는 동작을 설명하기 위한 도면이다.
- 도 4는 일 실시예에 있어서, 결과 변수의 인스턴스를 설명하기 위한 도면이다.
- 도 5는 일 실시예에 있어서, 속성에 대한 히스토그램을 설명하기 위한 도면이다.
- 도 6은 일 실시예에 있어서, 속성에 대한 박스플롯을 설명하기 위한 도면이다.
- 도 7은 일 실시예에 있어서, 상관계수 행렬을 설명하기 위한 도면이다.
- 도 8은 일 실시예에 있어서, 부스팅 알고리즘의 테스트 정확도를 설명하기 위한 도면이다.
- 도 9 및 도 10은 일 실시예에 있어서, 기타 측정 방법을 설명하기 위한 도면이다.
- 도 11은 일 실시예에 있어서, 예측을 위한 특징 중요도를 설명하기 위한 도면이다.
- 도 12 내지 도 14는 일 실시예에 있어서, 부스팅 알고리즘의 ROC 곡선을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0018] 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.
- [0020] 도 1은 일 실시예에 있어서, 심장질환 예측 시스템의 구성을 설명하기 위한 블록도이고, 도 2는 일 실시예에 있어서, 심장질환 예측 방법을 설명하기 위한 흐름도이다.

- [0021] 심장질환 예측 시스템(100)의 프로세서는 데이터 입력부(110) 및 심장질환 예측부(120)를 포함할 수 있다. 이러한 프로세서의 구성요소들은 심장질환 예측 시스템에 저장된 프로그램 코드가 제공하는 제어 명령에 따라 프로세서에 의해 수행되는 서로 다른 기능들(different functions)의 표현들일 수 있다. 프로세서 및 프로세서의 구성요소들은 도 2의 심장질환 예측 방법이 포함하는 단계들(210 내지 220)을 수행하도록 심장질환 예측 시스템을 제어할 수 있다. 이때, 프로세서 및 프로세서의 구성요소들은 메모리가 포함하는 운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다.
- [0022] 프로세서는 심장질환 예측 방법을 위한 프로그램의 파일에 저장된 프로그램 코드를 메모리에 로딩할 수 있다. 예를 들면, 심장질환 예측 시스템에서 프로그램이 실행되면, 프로세서는 운영체제의 제어에 따라 프로그램의 파일로부터 프로그램 코드를 메모리에 로딩하도록 심장질환 예측 시스템을 제어할 수 있다. 이때 데이터 입력부(110) 및 심장질환 예측부(120)의 각각은 메모리에 로딩된 프로그램 코드 중 대응하는 부분의 명령을 실행하여 이후 단계들(210 내지 220)을 실행하기 위한 프로세서의 서로 다른 기능적 표현들일 수 있다.
- [0023] 단계(210)에서 데이터 입력부(110)는 심장질환을 예측하기 위해 구성된 앙상블 모델에 심장질환 데이터를 입력받을 수 있다. 이때, 앙상블 모델에 입력되는 심장질환 데이터는 앙상블 모델의 학습을 위해 사용된 심장질환 데이터 셋과는 다른 새로운 데이터를 의미한다. 이러한 새로운 데이터에 나이, 성별, 심낭염, 휴식BP, 콜레스테롤, 단식 BP, 휴식심전도, 최대HR, 운동IA, 올드피크(Oldpeak), 슬로프(Slope), 형광 투시법으로 착색된 주요 혈관 수(Ca), 빈형 여부(Thal) 등의 속성 정보가 포함되어 있을 수 있다.
- [0024] 단계(220)에서 심장질환 예측부(120) 앙상블 모델을 이용하여 입력받은 심장질환 데이터로부터 심장질환을 예측할 수 있다. 예를 들면, 앙상블 모델은 심장질환 데이터 셋에 포함된 각 속성 정보를 이용하여 데이터 별(환자 별) 특징 정보가 학습되어 심장질환을 예측하도록 구축된 것일 수 있다. 심장질환 예측부(120)는 입력받은 심장질환 데이터로부터 심장질환 여부를 판단할 수 있다.
- [0025] 도 3은 일 실시예에 있어서, 심장질환을 예측하기 위해 앙상블 모델을 학습시키는 동작을 설명하기 위한 도면이다.
- [0026] 심장질환 예측 시스템은 심장질환을 예측하기 위한 데이터 셋을 이용하여 심장질환 예측을 위해 구성된 앙상블 모델을 학습시킬 수 있다. 자체적으로 앙상블 모델을 학습할 수도 있고, 외부의 앙상블 모델을 이용할 수도 있다. 실시예에서는 다양한 앙상블 학습 기법을 이용하여 앙상블 모델을 학습시키는 동작에 대하여 설명하기로 한다.
- [0027] 일례로, 공개적으로 사용 가능한 심장질환 데이터 셋을 웹 기반 Jupyter notebook(오픈 소스 플랫폼)에 가져올 수 있다. 필요한 라이브러리 패키지는 Sklearn에서 파이썬 프로그래밍 언어를 사용하여 설치할 수 있다.
- [0028] 처음에는 질병을 예측하기 위해 데이터 전처리 없이 부스팅 분류기가 적용될 수 있다. 탐색적 데이터 분석 후, 데이터의 전처리가 더 나은 결과를 획득하는데 중요한 역할을 할 수 있다는 것이 발견될 수 있다. 전처리 단계에서 결측값을 식별하고 결측값 보정 방법을 사용하여 다른 값으로 대체할 수 있다. 사분위간 범위 방법은 데이터 셋에 있는 이상치를 감지하고 대체할 수 있다. 일부 다른 필수 라이브러리는 데이터 셋에서 손상된 데이터를 확인하기 위해 실행될 수 있다. 데이터 셋은 70:30 비율로 나눌 수 있고, 여기서, 70%는 모델을 훈련시키는데 사용되고, 30%는 모델을 테스트하는데 사용될 수 있다. 결과를 검증하기 위해 k-fold 교차 검증(K=10)이 적용될 수 있다. 마지막으로, 원하는 결과를 획득하기 위해 데이터 전처리 후 세 가지의 부스팅 알고리즘이 적용될 수 있다.
- [0029] 앙상블 학습의 기술은 실제 문제를 해결하기 위해 거의 모든 분야에서 탐구된다. 이러한 모델은 다양한 질병의 더 나은 예측, 검출, 진단 및 예후에서 상당한 진전을 이루었다. 심장질환 예측 시스템은 심장질환 예측을 위해 복수 개(예를 들면, 3가지)의 앙상블 학습 기반 부스팅 알고리즘을 적용할 수 있다. 심장질환 예측 시스템은 앙상블 모델에 그래디언트 부스팅, AdaBoost 또는 XGBoost를 포함하는 하나 또는 둘 이상의 부스팅 알고리즘을 적용할 수 있다. 예를 들면, 심장질환 예측 시스템은 그래디언트 부스팅+AdaBoost, 그래디언트 부스팅+XGBoost, AdaBoost+XGBoost 또는 그래디언트 부스팅+AdaBoost+XGBoost를 이용하여 모델을 학습시킬 수 있다.
- [0030] 그래디언트 부스팅(Gradient boosting; GB): 약한 학습자는 순차적으로 훈련되며, 모든 추정기는 가중치를 적용하여 점진적으로 추가된다. 그래디언트 부스팅 알고리즘은 이전 추정기의 잔차 오차를 예측하는 데 초점을 맞추고 예측 값과 실제 값의 차이를 최소화하려고 시도한다.
- [0031] AdaBoost(AD): AdaBoost는 약한 학습자에 대한 사전 지식 없이 모든 가중치를 조정하여 작동한다. 기본 학습자

의 약점은 모델을 훈련하는 동안 추정기의 오류율로 측정된다. 의사 결정 트리 스템프는 분류 및 회귀 문제를 해결하기 위해 AdaBoost 알고리즘과 함께 널리 사용된다.

[0032] XGBoost(XGB): XGBoost는 서로 다른 종류의 의사결정 트리(약한 학습자)를 결합하여 유사성 점수를 독립적으로 계산함으로써 작동한다. 그레디언트 하강 및 정규화 프로세스를 적용하여 훈련 단계에서 과적합 문제를 극복하는 데 도움이 된다.

[0033] 예를 들면, 심장질환 예측 시스템은 앙상블 모델을 이용하여 심장질환을 예측하기 위해 캘리포니아 대학교 어바인(UCI)의 머신러닝 저장소에서 공개적으로 사용할 수 있는 심장질환 데이터 셋을 사용할 수 있다. 데이터 셋은 광범위한 인구 통계를 포함하여 심장질환과 관련된 임상 특징이 풍부하다.

[0034] 이러한 심장질환 데이터 셋에는 1329개의 인스턴스와 14개의 속성이 포함되어 있으며, 여기서 처음 13개의 속성은 술어/독립 변수이고, 마지막 속성은 종속/대상 변수이다. 표 1에는 고려된 속성, 속성 설명, 해당 측정값, 범위 값에 대한 정보가 나와있다. 표 1에 나이, 성별, 심낭염, 휴식BP, 콜레스테롤, 단식 BP, 휴식심전도, 최대HR, 운동IA, 올드피크(Oldpeak), 슬로프(Slope), 형광 투시법으로 착색된 주요 혈관 수(Ca), 빈형 여부(Thal), 결과(클래스 속성) 등에 대한 정보가 설명되어 있다.

[0035] 표 1:

Attribute	Description	Measurement	Value range
Age	Age of an individual	Years	29 to 77
Sex	Gender of an individual	1 = male, 0 = female	0 or 1
Cpericarditis	Degree of chest pain	Low, moderate, high, extremely high	0 to 3
RestingBP	Blood pressure of an individual while at rest (inactive)	Hg level (in mm)	94 to 200
Cholesterol	Level of serum cholesterol	mg/dl	126 to 256
FastingBP	Glucose level in an empty stomach (fasting)	Greater than 120 mg/dl (1 = true, 0 = false)	0 or 1
RestingECG	Resting electrocardiographic results of an individual while inactive	0 = normal, 1 = having ST	0 to 2
MaximumHR	Highest heart rate recorded.	Beats per minute	71 to 202
ExercisefA	Doing exercise with angina disease	1 = yes, 0 = no	0 or 1
Oldpeak	ST depression, caused by doing exercise in comparison to being inactive	Numeric value	Relative
Slope	Slope of the old peak value in the ST segment while an individual is doing an exercise	0 = downsloping, 1 = flat, 2 = upsloping	0 to 2
Ca	No. of major vessels colored by fluoroscopy	Numeric	0 to 3
Thal	Whether an individual has thalassemia or not	3 = normal, 6 = fixed defect, 7 = reversible defect	3 to 7
Outcome	Class attribute	0 = no heart disease, 1 = heart disease	0 or 1

[0036]

[0037] 데이터 셋은 기술 통계(descriptive statistics) 데이터 특성을 식별하는 데 중요한 역할을 한다. 이는 데이터를 요약하여 인간이 해석하기 쉽게 한다. 표 2는 기록 카운트, 최소(min) 값, 최대(max) 값, 평균 및 표준 편차(Std)와 같은 측정값과 함께 임상 속성의 통계적 측정을 설명한다. 예를 들어, 연령 속성은 평균 값으로 54.41, 표준 편차로 9.07이며, 최대 및 최소 연령 수는 각각 77세와 29세이다. 이러한 통계적 측정은 나머지 속성에 대해서도 계산될 수 있다.

[0038] 표 2:

Attributes	Count	Mean	Std	Min	Max
Age	1328	54.41	9.07	29	77
Sex	1328	0.69	0.46	0	1
Cpericarditis	1328	0.94	1.02	0	3
RestingBP	1328	131.61	17.51	94	200
Cholesterol	1328	246.06	51.62	126	564
FastingBP	1328	0.14	0.35	0	1
RestingECG	1328	0.52	0.52	0	2
MaximumHR	1328	149.23	22.97	71	202
ExerciselA	1328	0.33	0.47	0	1
Oldpeak	1328	1.06	1.17	0	1
Slope	1328	1.38	0.61	0	2
Ca	1328	0.74	1.02	0	4
Thal	1328	2.32	0.61	0	3
Outcome	1328	0.52	0.49	0	1

[0039]

[0040] 도 4는 일 실시예에 있어서, 결과 변수의 인스턴스를 설명하기 위한 도면이다.

[0041] 기계/양상불 학습 모델은 사용된 데이터 셋이 문제 기술문에 대해 균형을 이루지 못할 경우 좋지 않은 결과를 제공한다. 일부 상황에서는 대상 클래스가 균등하게 분포되어 있지 않으면 일부 샘플링 기술을 사용하여 균형 잡힌 데이터셋을 만들 수 있다. 심장질환 데이터 셋에는 도 4와 같이 클래스 1은 심장 질환 있음(691 인스턴스)이고, 클래스 0은 심장질환이 없음(637 인스턴스)인 클래스가 잘 혼합되어 있다.

[0042] 도 5는 일 실시예에 있어서, 속성에 대한 히스토그램을 설명하기 위한 도면이다.

[0043] 히스토그램은 데이터 샘플의 분포를 시각화하고 해석하는 데 사용된다. 히스토그램은 균일, 정규, 왼쪽 비대칭 및 오른쪽 비대칭으로 표시될 수 있다. 도 5는 값 범위 내의 모든 속성을 그룹화하는 정규 분포 히스토그램을 보여준다. x축은 속성의 특성을 나타내고 y축은 해당 속성의 값을 나타낸다.

[0044] 도 6은 일 실시예에 있어서, 속성에 대한 박스플롯을 설명하기 위한 도면이다.

[0045] 도 6은 데이터 셋에 존재하는 각 속성의 박스플롯을 나타낸 것으로, 속성에 대한 박스플롯을 나타내기 위해 확률 밀도 함수를 사용하는 사분위간 범위 방법은 데이터 셋의 이상치를 처리하는데 사용될 수 있다. 예를 들면, 공복 혈압에서는 단일 이상치가 검출된 반면, 휴식기 혈압에서는 다중 이상치가 검출되어 z-점수법으로 대체될 수 있다.

[0046] 도 7은 일 실시예에 있어서, 상관계수 행렬을 설명하기 위한 도면이다.

[0047] 상관계수 분석(The correlation coefficient analysis; CCA) 방법은 데이터 셋 속성 간의 관계를 식별하고 도표화하는 데 사용될 수 있다. 독립적 속성과 종속적 속성 셋 사이에 강력한 연관성/관계가 존재하는 경우 데이터 셋이 양호한 것으로 간주될 수 있다. 도 7은 질병 예측에 사용되는 모든 속성의 상관계수 분석을 나타내며, x축과 y축 내에서 +1에서 -1 사이의 관계 범위가 존재한다. 셀 값은 교차하는 속성 간의 관계 정도를 나타낸다. 예를 들어, 휴식기 혈압과 나이 사이의 관계 값은 0.12이다.

[0048] 실시예에서는 심장질환 예측을 위한 부스팅 알고리즘을 사용하여 달성한 실험 세부 사항과 결과에 대하여 설명하기로 한다. 이 후, 제안된 프레임워크를 구현한 후의 모든 결과가 체계적으로 표시되고 분석될 수 있다. 결과는 질병 예측을 위한 전처리 전과 전처리 후의 두 가지 모듈로 제공될 수 있다. 평가는 정밀도, 재현율, f1-점수, ROC(receiver operation curve) 곡선 및 고려되는 부스팅 알고리즘의 이동 시간과 같은 성능 평가 메트릭 측면에서 광범위하게 논의될 수 있다.

[0049] 데이터 전처리는 머신러닝 방법을 모델에 적용하기 전에 강력하고 신뢰할 수 있는 시스템을 개발하는 데 필수적이다. 실시예에서는 결측값을 식별하고 결측값 보정 방법으로 대체할 수 있다. 처음에는 isnull() 메소드를 사용하여 모든 결측값을 검출한 다음 SimpleImputer() 메소드로 평균 및 모드 대체 기법을 실행하여 결측값을 채울 수 있다. 이러한 프로세스는 열의 평균, 중간, 최빈값을 사용하여 모든 결측값을 대체할 수 있다. 모든 데이터 샘플의 분포를 이동하고 평균을 0으로 만들기 위해 Z-점수 기법을 사용한 사분위간 범위 방법을 사용하

여 이상치를 검출하고 대체할 수 있다.

- [0050] 하드웨어/소프트웨어 사양 및 계산 시간에 대하여 설명하기로 한다. 연구 작업을 수행하기 위해 HP Z60 워크스태이션이 사용될 수 있다. 시스템의 하드웨어 사양은 Intel XEON 2.4GHz CPU(12 코어), 4GB RAM, 1TB 하드 디스크 및 Windows 10 Pro-64비트가 사용될 수 있다. 이 시스템에서 알고리즘 ADB, XGB, GB를 실행하는 데 각각 4.23초, 3.57초, 4.51초가 소요될 수 있다. 하드웨어 구성 요소를 제외하고 구현에 사용되는 소프트웨어는 그래픽 사용자 인터페이스 기반 아나콘다 네비게이터, 웹 기반 컴퓨팅 플랫폼 주피터 노트북, 프로그래밍 언어로서의 파이썬이다.
- [0051] 도 8은 일 실시예에 있어서, 부스팅 알고리즘의 테스트 정확도를 설명하기 위한 도면이다.
- [0052] 분류기의 정확도를 판단하기 위해 XGBoost(XGB), AdaBoost(ADB), 그레디언트 부스팅(GB)을 포함하는 부스팅 알고리즘이 사용될 수 있다. 전처리 기법을 적용하지 않으면 XGB, ADB, GB와 같은 분류기의 정확도는 각각 87.50%, 81.50%, 86%이다. 전처리 기법을 적용한 결과 그레디언트 부스팅이 92.20%로 가장 높은 정확도를 획득하여, 다른 부스팅 알고리즘을 능가했고, AGB와 ADB가 89.61%로 뒤를 이었다.
- [0053] 도 9 및 도 10은 일 실시예에 있어서, 기타 측정 방법을 설명하기 위한 도면이다.
- [0054] 세 분류기의 정밀도, 리콜 및 f1-점수는 데이터 전처리 전후에 계산될 수 있다. 도 9와 도 10에 나타난 바와 같이 두 클래스(0: 심장병 없음, 1: 심장병) 모두에 대해 값이 계산(백분율로)될 수 있다. XGB는 전처리 없이 모든 측정에 대해 가장 잘 수행된 반면 ADB는 가장 나쁜 수행을 수행했다. 전처리를 통해 GB는 대부분의 측정에서 가장 우수한 성능을 발휘한 반면, XGB와 ADB는 거의 동일한 결과가 달성되었다.
- [0055] 도 11은 일 실시예에 있어서, 예측을 위한 특징 중요도를 설명하기 위한 도면이다.
- [0056] 특징 중요도는 출력 특징(종속/대상 변수)을 예측하는 기여도를 기반으로 입력 특징(독립적인 술어 변수)의 점수를 계산하는 프로세스이다. 예측 결과를 개선하기 위해 기계/양상불 학습 모델을 개발하는 데 중요한 역할을 한다. 실시예에서 특징 중요도 점수(F-점수)는 훈련 과정에서 속성이 분할에 사용되는 횟수를 나타낸다. 특징(예: 콜레스테롤)의 F-점수가 높으면 그것이 중요한 속성임을 나타낸다. 도 11은 F-점수를 기준으로 내림차순으로 예측에 대한 모든 속성의 기여도를 보여준다. 예를 들어, 콜레스테롤은 예측에서 가장 높은 의미를 갖는 반면, 공복 혈압은 가장 낮다.
- [0057] 도 12 내지 도 14는 일 실시예에 있어서, 부스팅 알고리즘의 ROC 곡선을 설명하기 위한 도면이다.
- [0058] ROC 곡선은 서로 다른 임계값에서 부스팅 알고리즘을 고려하는 예측 능력을 보여주는 데 사용될 수 있다. 각각 x축과 y축을 따라 false-positive 비율 대 true positive 비율을 나타낸다. ROC 곡선을 사용하여 양상불 모델이 클래스(0-심장병 없음 및 1-심장병)를 얼마나 잘 구분하는지 분석될 수 있다. ROC 곡선이 높으면 양상불 모델이 0과 1 사이의 양호한 결과를 예측하고 있음을 의미한다. 양상불 모델의 AUC가 1 근처에 있으면 분리성 측도가 우수하다는 것을 의미하며, AUC가 0 근처에 있으면 분리성 측도가 최악임을 의미한다. AUC 값이 0.5이면 모델이 클래스를 효과적으로 분리하는 데 작동하지 않는다. 도 12를 참고하면, ADB의 ROC 곡선, 도 13을 참고하면, XGB의 ROC 곡선, 도 14를 참고하면, GB의 ROC 곡선을 나타낸 것이다. 수치로부터, XGB가 가장 잘 수행되고, 그 다음이 GB와 ADB라고 결론을 도출할 수 있다.
- [0059] 실시예에서 제안된 심장질환 예측 방법은 심장 질환 예측에 대한 다양한 평가 지표 측면에서 좋은 결과가 획득될 수 있다. 실시예에서 제안된 심장질환 예측 방법의 성능은 기술, 데이터 셋 및 정확도 측면에서 여러 관련 연구와 비교될 수 있다. 실시예에서 제안된 심장질환 예측 방법이 다양한 평가 지표, 특히 심장 질환 예측의 정확성 측면에서 좋은 결과를 획득할 수 있다. 결측값 처리를 위한 결측값 보정, 박스 플롯 방법을 사용한 이상치 검출 및 보정과 같은 기법을 사용하여 다른 관련 작업보다 정확한 결과를 획득할 수 있다.
- [0060] 심장질환을 효과적으로 예측하기 위해 부스팅 알고리즘이 적용될 수 있다. 데이터셋의 예측 결과와 품질 평가를 개선하기 위해 보정, Z 점수, 클리닝 방법과 같은 다양한 전처리 방법이 사용될 수 있다. 실시예에서는 또한 전처리 기법 적용 전후에 XGBoost, AdaBoost, Gradient Boost의 세 가지 부스팅 알고리즘을 실행할 수 있다. 실험 결과는 다른 통계/ML 측정을 사용하여 평가될 수 있다. 실험 결과 그레디언트 부스팅이 92.20%의 가장 높은 정확도를 달성한 것으로 나타났다. 그레디언트 부스팅 알고리즘은 정밀도, 리콜 및 f1 점수와 같은 다른 지표에 대해서도 보다 정확한 결과를 달성하였다.
- [0061] 마지막으로, 특징 중요도 프로세스를 사용하여 최종 결과에 대한 독립적 특징의 기여도가 계산될 수 있다. bagging 및 stacking과 같은 다른 양상불 학습 기술을 사용하여 심장질환 예측의 효과를 향상시킬 수도 있다.

실시예에서 제안된 심장질환을 예측하는 방법은 연구 작업의 범위를 확장하기 위해 특징의 공통성을 공유하는 다른 의료 데이터셋에 사용될 수 있다. 심혈관 질환을 더 잘 검출하고 예측하기 위해 딥 러닝 기술을 탐구할 수도 있다.

[0062] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(configuration)도 가능하다.

[0063] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0064] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0065] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

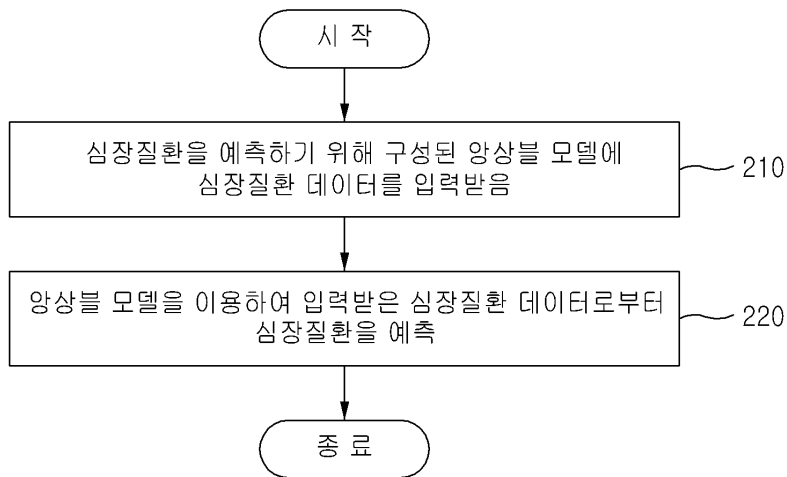
[0066] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

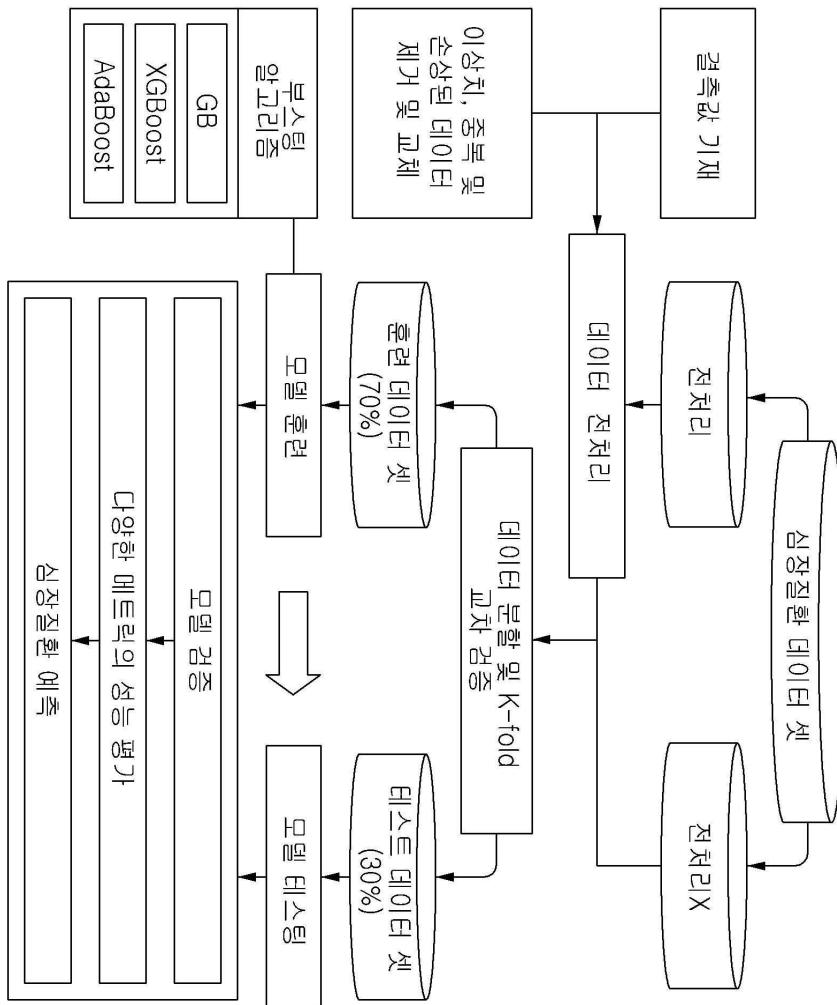
도면1



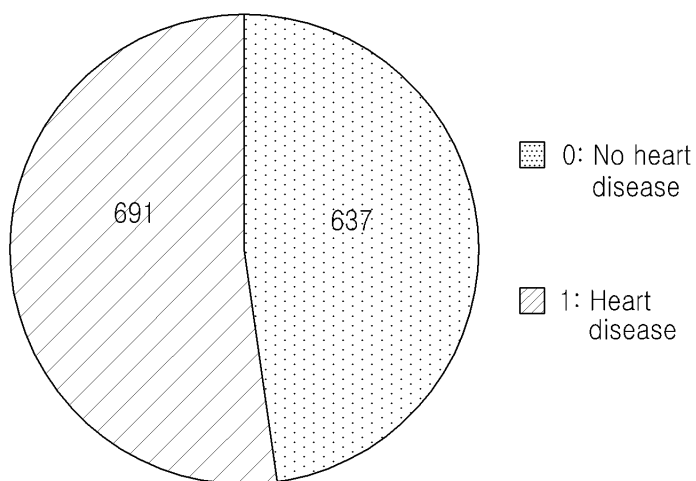
도면2



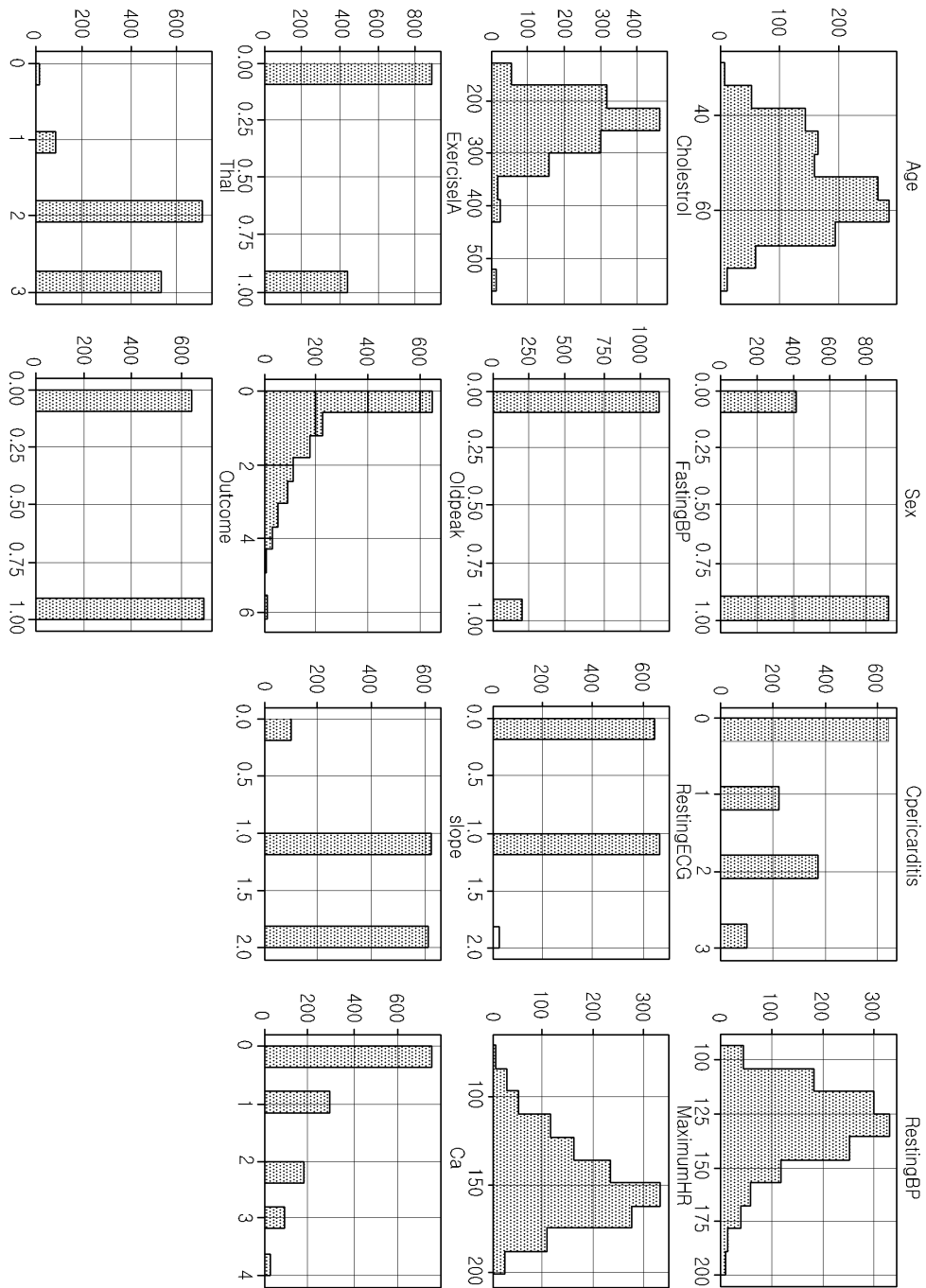
도면3



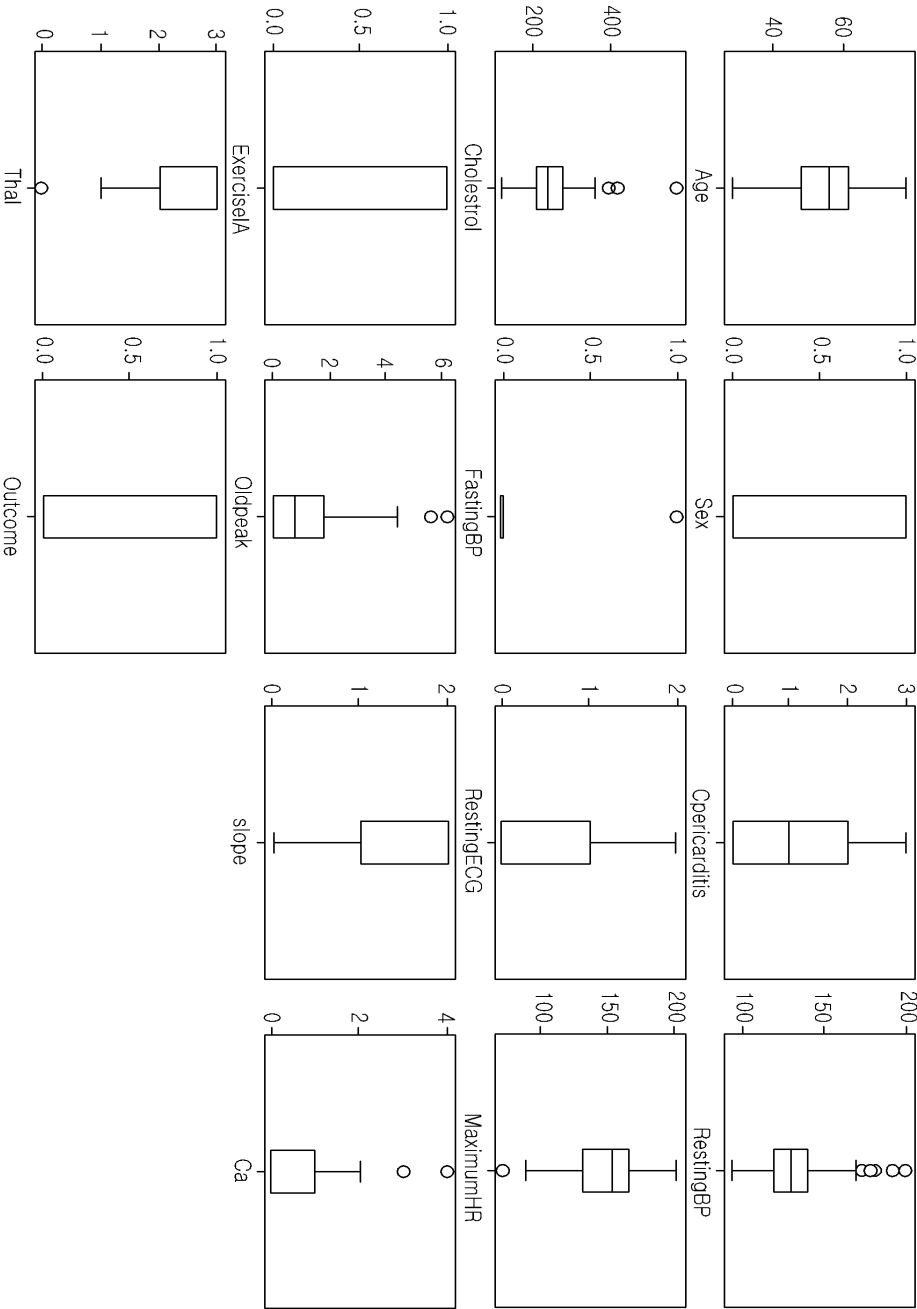
도면4



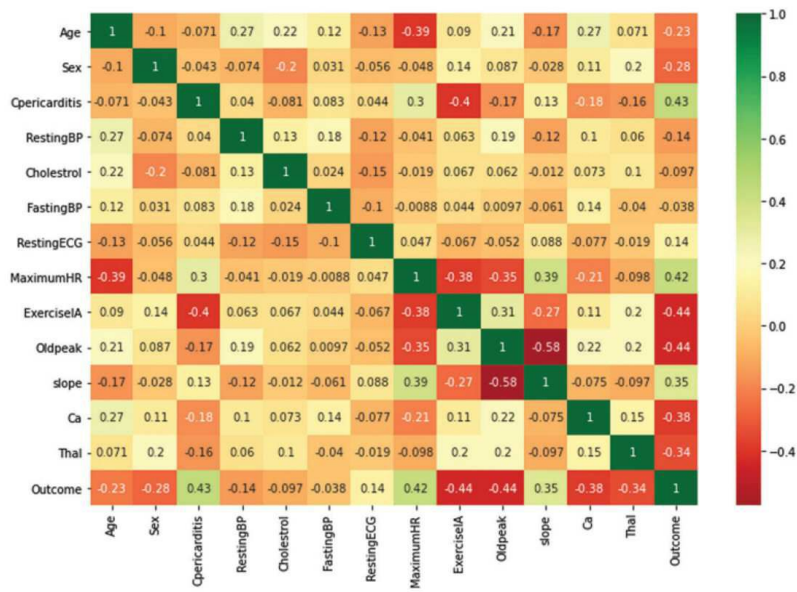
도면5



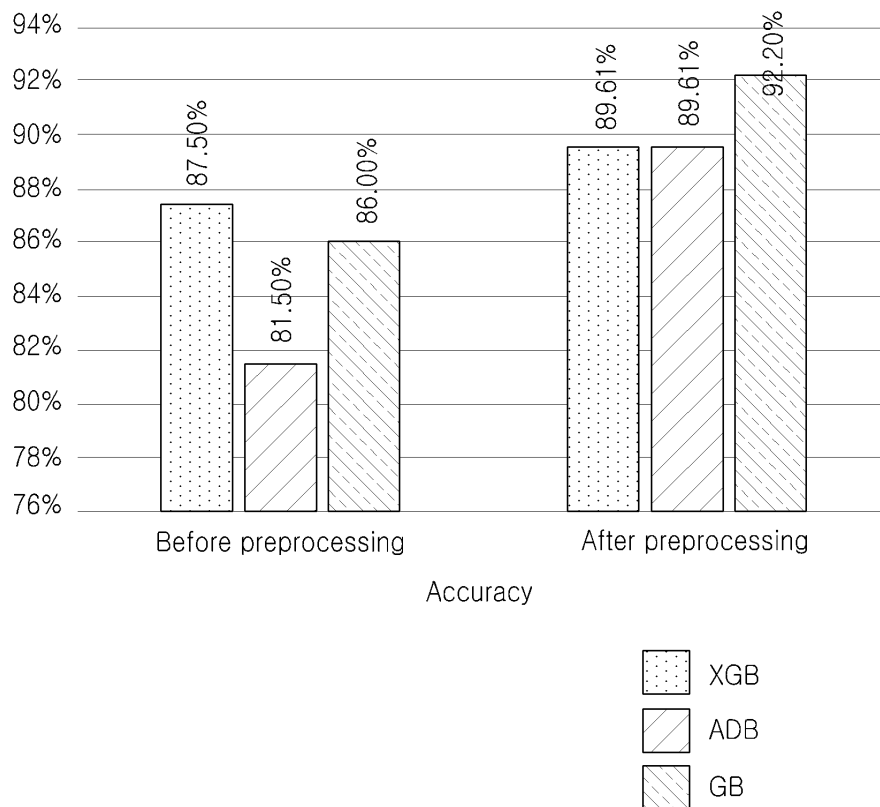
도면6



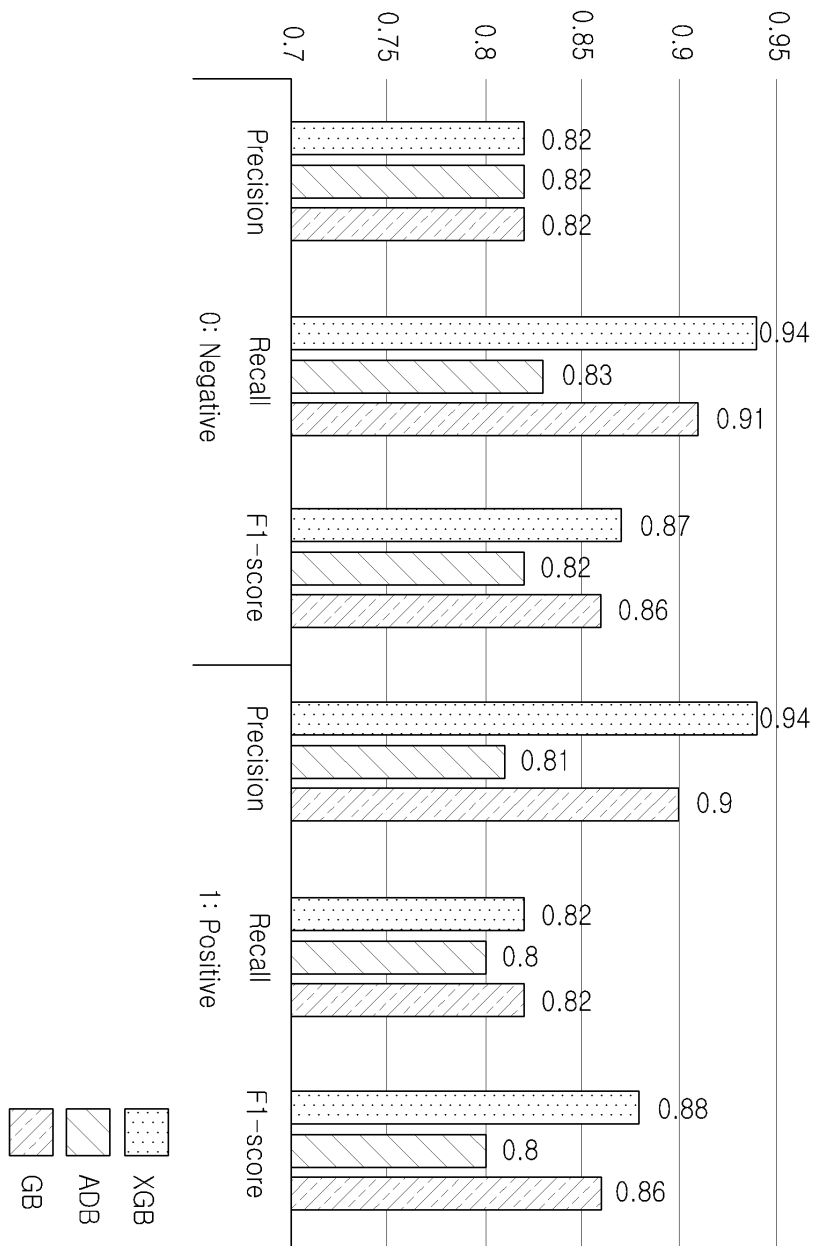
도면7



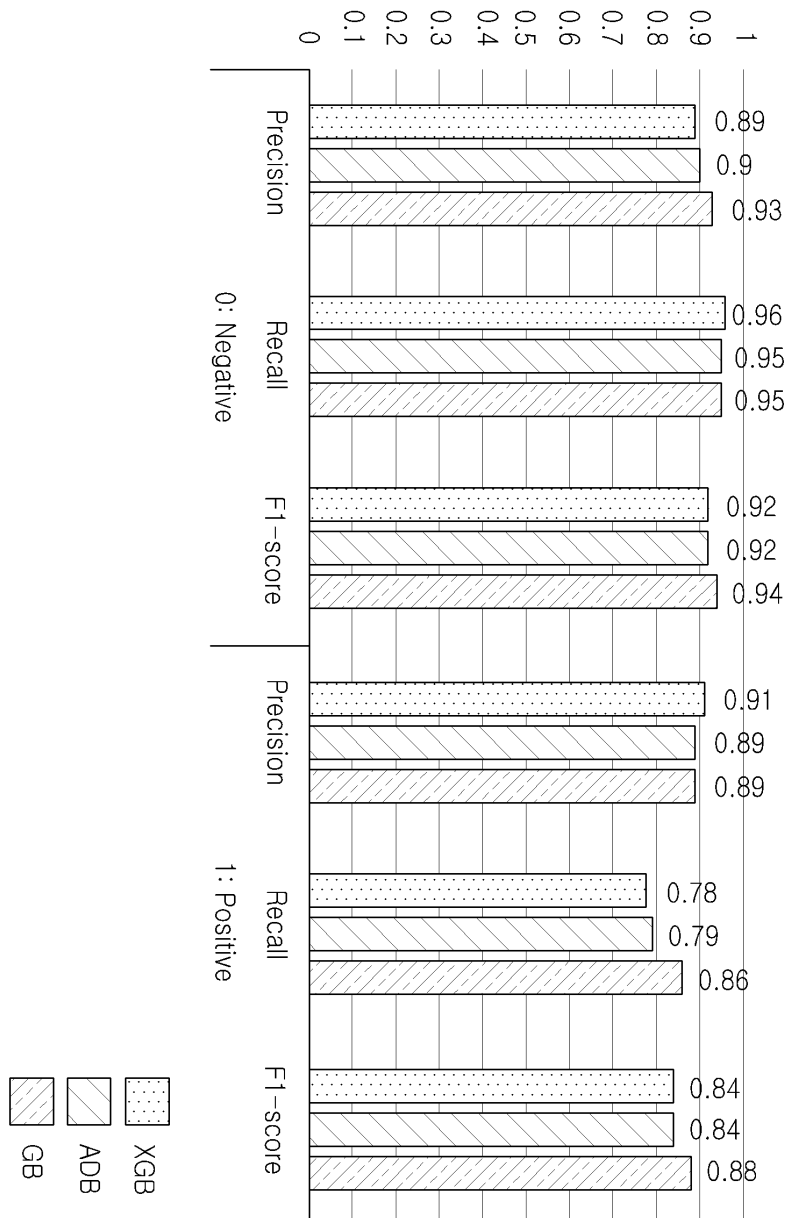
도면8



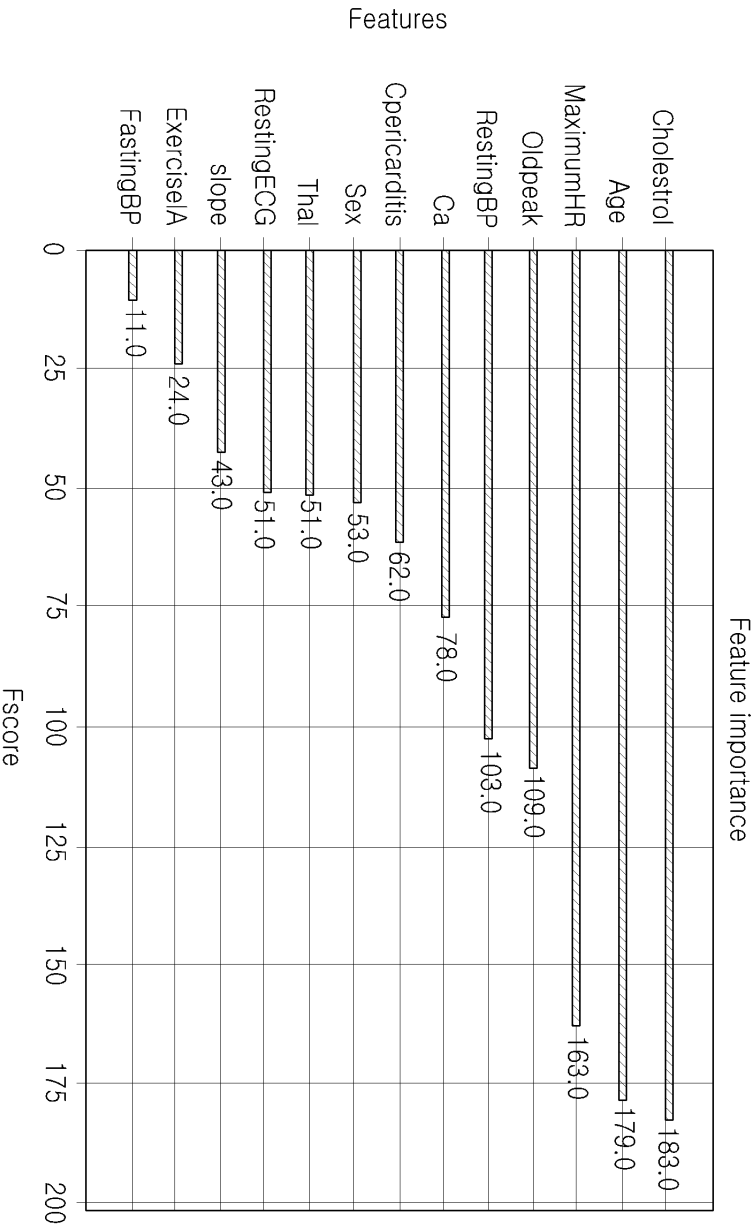
도면9



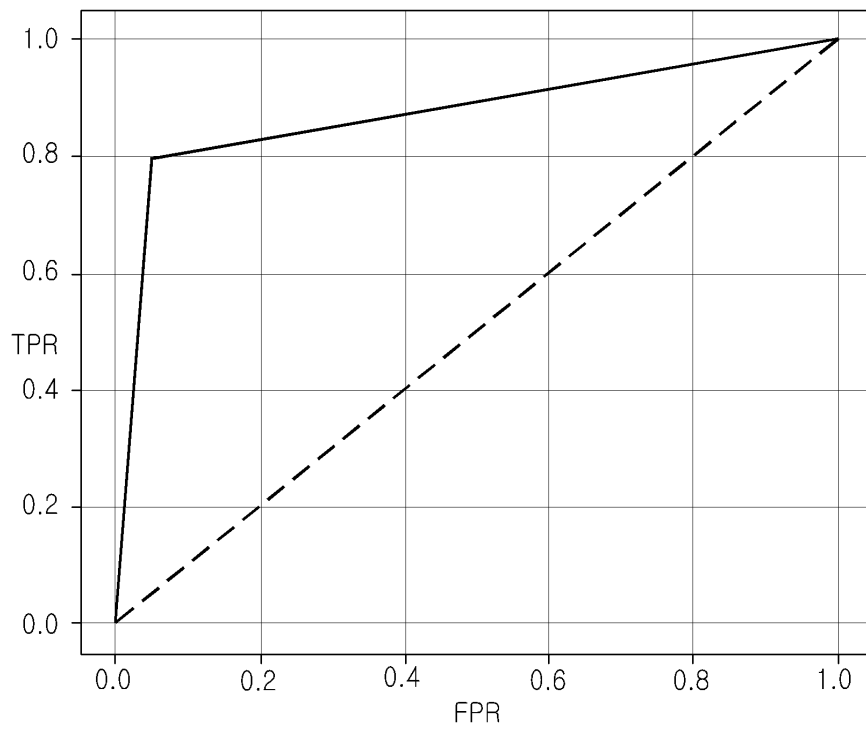
도면10



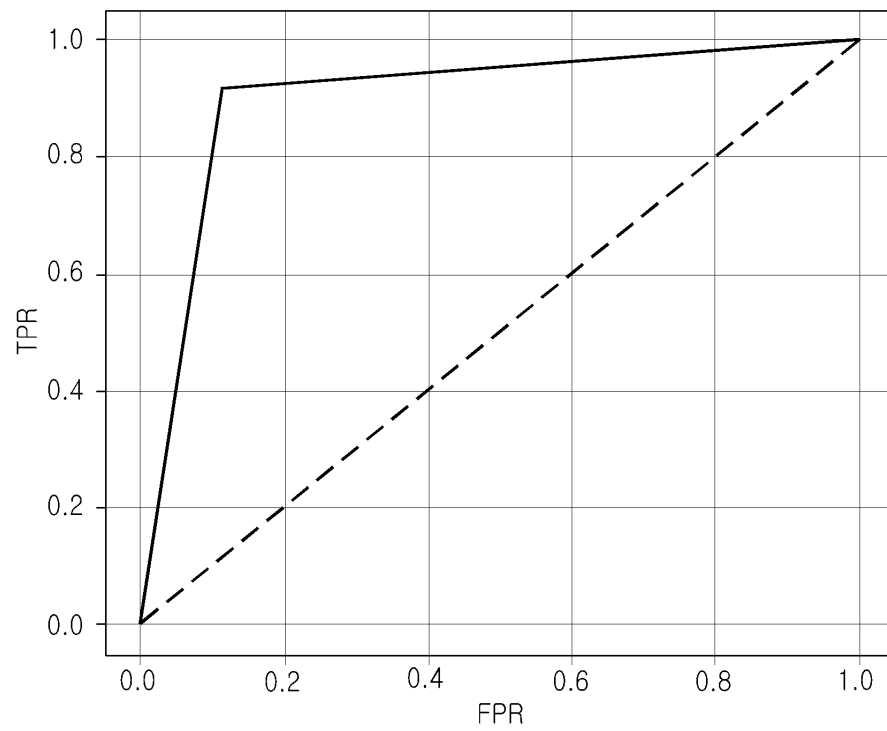
도면11



도면12



도면13



도면14

