

2024 년도 기초과학융합연구소 학부생 동계 인턴십 신청서

성 명	이건	소 속	정보통계보험수리학과	*타 과제참여 여부(0/X)	X
학 번	██████████	이메일	██████████	연 락 처	██████████

○연구제목

다양한 머신러닝 알고리즘을 활용한 당뇨병 예측 모델 개발 및 성능 분석 최적의: 모델 선정 및 구현을 위한 종합적 접근 ○연구배경

저는 2024-2 학기 7+1 비교과 활동의 일환으로 두드림 프로젝트를 진행한 경험이 있습니다 해당. 프로젝트의 주제는 ‘당뇨병 예측 모델 개발 및 성능 분석으로 이번’, 연구 주제와 동일한 분야를 다루고 있습니다 당시에는. 로지스틱 회귀 모델을 사용하여 연구를 진행했으며, k-fold 검증 방식을 활용해 양성 레이블에 최적의 가중치를 부여한 후 모델의 성능을 평가했습니다 결과적으로. AUC 는 0.76, 민감도는 0.62, 특이도는 0.75 로 산출되었지만 개인적으로는, 성능 지표에 아쉬움을 느꼈습니다 이. 연구를 바탕으로 노모그램까지 생성했으나 성능을, 개선할 여지가 많다고 판단했습니다. 데이터는 2023 년 국민건강영양조사 데이터를 활용했지만 데이터, 수가 모델 훈련에 부족하다고 생각하여 2023 년 년~2018 데이터를 통합해 분석을 시도했습니다 그러나. 통합 과정에서 다수의 설명변수가 손실되고 결측값, 제거 과정에서 상당량의 데이터가 손실되어 모델의 성능 저하를 겪게 되었습니다. 이후 2024-2 학기에 수강한 ██████████ 교수님의 데이터 마이닝 수업을 통해 여러 머신러닝 기반 분류 알고리즘에 대해 학습할 수 있었습니다 당시. 진행했던 프로젝트에 배깅(Bagging), 부스팅(Boosting), 랜덤 포레스트(Random Forest)와 같은 앙상블 기법을 활용하면 데이터 부족 문제를 어느 정도 극복할 수 있으리라 판단하였습니다. 제가 본 연구 주제를 개인적으로 진행하기보다는 정보통계보험수리학과, ██████████ 교수님이 이끄시는 연구실에서 진행하고 싶었던 이유는 의료 데이터를 다룰 때 잘못된 모델 결과가 환자에게 부적절한 영향을 미칠 수 있는 점을 깊이 우려했기 때문입니다 이를. 해결하기 위해 전문적인 교수님의 지도를 받아 정확도 높은 모델을 설계하고 싶었습니다 또한 수업. , 시간에 배운 다양한 모델을 실제 연구에 적용해보고자 하는 의지도 있었습니다 아울러. , 2023 년 국민건강영양조사 데이터 외에 교수님 연구실에서 보유한 추가적인 데이터를 활용할 기회가 있을 것이라 기대했기 때문에 이번, 기초과학융합연구소에서 개설한 동계 인턴십에 지원하게 되었습니다.

○ 연구의 중요성

개인적 중요성

동계인턴십에서 진행하고자 하는 저의 연구는 제가 데이터 마이닝 수업에서 학습한 머신러닝 기법을 실제 데이터에 적용할 수 있는 소중한 기회라는 점에서 매우 중요한 의미를 가집니다 특히 배깅. , (Bagging), 부스팅(Boosting), 랜덤 포레스트(Random Forest) 등 다양한 앙상블 기법을 활용함으로써, 이론적으로만 접했던 머신러닝 기법을 실제 의료 데이터를 통해 구현하고 성능을 분석할 수 있습니다 이. 는 단순한 학습을 넘어 실제 저의 역량을 강화하고 머신러닝, 기반 데이터 분석에서 실질적이고 전문적인 경험을 축적할 수 있는 중요한 발판이 될 것입니다 또한 제가. , 이전에 진행했던 프로젝트에서 느낀 한계를 극복하며 분석, 능력을 심화하고 보다 정교한 모델링을 설계할 수 있는 기회를 제공한다는 점에서 연구 참여는 개인적으로도 큰 의미를 지닙니다.

○연구계획

다음 페이지에
기술하겠습니다.

○연구계획

1. 설명변수 설정

- 목적: 당뇨병 발생에 영향을 미칠 가능성이 높은 요인을 가설로 설정.
- 예 나이 체질량지수: , (BMI), 혈압 심혈관, 질환 이력 생활, 습관흡연 음주 혈액(,), 검사 결과콜레스테롤(, 혈당 호르몬), 수치 약물, 복용 여부 등.
- 심리적 요인과 생활적 요인(예 스트레스: 지수 신체, 활동 수준도) 포함하여 모델의 예측력을 높임. 설명변수를 과도하게 설정하면 다중공선성문제가 발생할 수 있으므로 이후, 단계에서 이를 조정함.

여러 당뇨병 예측 모델 논문을 참조하여 설명변수 후보군을 선정하고 카이제곱을 통한 분포 별 차이를 확인하여 설명변수에 대한 후보군을 산정하고자 합니다.

2. 설명변수 데이터 수집

- 데이터를 수집하기 위해 국민건강영양조사(KNHANES)와 같은 공신력 있는 공공 데이터베이스 활용. • 결측값 처리:
- 결측값이 적을 경우 평균또는 중앙값 대체. 결측값이 많을 경우 데이터 보간(interpolation) 또는 결측값 기반 예측 모델 적용. 가능하다면 교수님 랩실에서 제공받을 수 있는 데이터를 사용하기를 희망합니다.

3. 정규성 분석 및 데이터 전처리.

설명변수들이 정규분포를 따르는지 확인:

- Box Plot, Normal Q-Q Plot 등을 활용해 변수 분포 시각화.
 - 정규분포를 만족하지 않으면 로그 변환 루트, 변환등의 정규화 기법 적용. 이상치 처리:
 - 이상치를 탐지하기 위해 IQR(사분위 범위), Z-Score 등 사용.
 - 이상치 제거 여부는 분석 목적에 따라 결정:
 - 의료 데이터의 경우 이상치가, 의미 있는 경우도 있으므로 자동 제거보다는 의학적 판단 병행. 변수 변환: 범주형 변수는 One-Hot Encoding 또는 Label Encoding 으로 변환. 단위 통일예 혈압(: mmHg, BMI kg/m²). 스케일링: 해당 모델에 맞는 스케일링을 진행 할 예정입니다.
- 로지스틱 회귀와 같은 알고리즘은 정규화(Normalization)할 것이고 트리 기반 모델(Random Forest, XGBoost)은 스케일링을 고려하지 않을 생각합니다. 즉 여러 머신러닝 모델에 맞는 데이터 프레임을 여러개 생성 할 예정입니다.

4. 데이터 분석 1. 설명변수와

반응변수의 영향도 분석:

- 변수별로 카이제곱 검정 상관분석 검정, t- 을 수행하여 변수의 중요도 확인. 설명변수 간 상관관계 분석:
- 피어슨 상관계수로 변수 간 관계를 시각화.
- 높은 상관관계(> 0.8)는 다중공선성 문제로 이어질 수 있음. 다중공선성 판단:

****VIF(Variance Inflation Factor)****로 다중공선성을 확인.

높은 VIF 를 가진 변수는 제거하거나 변수 결합 등으로 해결.

간결성 원칙:

불필요한 설명변수를 제거하여 모델의 복잡성을 줄이고 일반화, 능력을 높임. 설명변수 20 개 내의 모델을 만들고 이를 간결성의 원칙을 고려한 설명변수가 8 개인 모델을 만들고 성능 지표를 비교 후 모델을 고려하고자 합니다고려방식. (deviance 등)

5. 예측 모델 개발

1. 후보 모델 선정:

- 다양한 머신러닝 분류 알고리즘 활용:
- 로지스틱 회귀(Logistic Regression): 해석 용이성 본과. - 회귀분석 수업
- 서포트 벡터 머신(SVM): 고차원 데이터 처리 강점 본과. - 데이터 마이닝 수업
- 랜덤 포레스트(Random Forest): 변수 중요도 해석 용이 과적합, 방지 본과. - 데이터마이닝 수업
- XGBoost, LightGBM: 부스팅 기반 모델로 예측 정확도 우수.
- k-Nearest Neighbors (k-NN): 단순하지만 데이터 구조에 민감 본과. - 데이터마이닝 수업 • 딥러닝 모델(Neural Networks): 대규모 데이터셋에서 활용 가능. 모델 성능 평가:
- k-fold 교차 검증:
- 데이터를 k 개로 분할하여 모델을 훈련하고 테스트.
- 데이터 부족 문제를 해결하고 모델의 일반화 성능을 확인. 주요 성능 지표:

AUC(ROC 곡선 하단 면적): 모델의 분류 성능 평가.

정밀도(Precision), 민감도(Sensitivity): 양성 레이블 분류 성능. F1

점수: 정밀도와 민감도의 조화 평균.

기법 활용:

(Bagging):

포레스트와 같은 배깅 기반 모델은 훈련 데이터를 여러 샘플로 분할하여 각각 독립적인 모델 훈련하여 부족 문제 해결 및 모델 안정성 증가. 부스팅(Boosting):

XGBoost, LightGBM 은 부스팅 기법의 대표적 모델로 이전, 모델의 오차를 보완하며 학습.

적은 데이터로도 높은 정확도를 달성 가능.

6.모델 평가 및 배포

1. 최적 모델 선정:

- R^2 , Adjusted R^2 , AIC/BIC, 쿡의 거리 통계량 등을 고려하여 최적의 모델 선택. Test Set 성능 평가:
- 학습 데이터와 독립적인 테스트 세트를 사용해 일반화 성능 확인.

API 개발 및 배포:

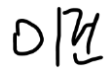
모델을 Flask, FastAPI 와 같은 웹 프레임워크를 사용해 API 로 구현.

Docker 와 같은 컨테이너 기반 기술을 사용해 배포하여 확장성과 유지보수성 확보. 사용자 입력 데이터를 받아 예측값을 반환하는 시스템 설계.

API 개발 및 배포는 개인 vlog 에 개시 할 예정이며 교수님 승인 하에 개시 할 예정입니다.

위와 같이 기초과학융합연구소 학부생 연구 인턴십 프로그램에 지원합니다.

2024. 12. 16..

지원자: 이건 (/인서명) 

지도교수 :  (/인서명) 