

Portfolio

2020.10 ~ 2022.09

성균관대학교 삼성융합의과학원

스마트헬스랩

송이준

목차

1. 신장 기증환자 예후 예측 (졸업논문)
2. 다국가 기반 손상환자의 뇌내출혈 예측
3. 응급실 처치 추천모델의 효용성 평가
4. CPR dashboard 개발

1. 신장 기증환자 예후 예측

1) 연구목표

- 신장 기증자의 기증 수술 후의 eGFR 예측 (regression)

2) 기여도

- 현재 논문 작성중 (1저자)
- 동일한 도메인의 기존 연구 리서치
- 실험 결과에 대한 ppt 준비 및 발표
- 데이터 전처리, 머신러닝 기반의 데이터마이닝 분석 결과 제시
- 테스트용 웹 어플리케이션 개발 (javascript, python)

3) Skills

- python, scikit-learn, pandas, numpy, javascript, react, flask, AWS

3) 분석과정 요약

a) EDA

- 변수 분류 (categorical, numeric)
- Basic demographic 확인: 데이터의 변수별 최대값, 최소값, 평균, 중간값 등 확인 (데이터 시각화)

b) Data Preprocess

- exclusion criteria 적용: 사용 할 수 없다고 판단되는 데이터 rows 제거 (환자 기본정보, 필수변수 누락된 경우 등..)
- 결측값 대체 (극단값 개별적 확인 및 수정, 또는 문제없는 변수의 경우 일괄 대표값으로 대체)
- 모델링에 사용하기 위한 최종 데이터 생성 : demographic 최종 확인

c) Modeling

- 머신러닝 모델 트레이닝
- 모델 튜닝: 모델별 하이퍼파라미터 튜닝

d) Result summary

- 성능지표 도출: 기존 연구 논문들에 사용된 성능지표 리서치
- 성능지표 제시 (Table, Plot)

1. 신장 기증환자 예후 예측

Study design & Population

목표 : 신장 기증자의 기증 후 예후 (eGFR) 예측모델 개발

Title : Predicting the post eGFR of Kidney Donors using Machine Learning Techniques with conventional regression models

- 신장내과와 함께 연구 진행중 (전준석 교수님, 장혜련 교수님)
- 삼성서울병원 EMR 신장 기증자 데이터 (2009 . 07 - 2020 . 12) : 823건
- 모든 연령 포함
- Input : 총 31개 (수술 이전의 여러가지 검사 결과값 포함)
- Outcome : 12개월 차의 eGFR
(신장 기증 수술 이후, 3, 6, 12 개월 차의 eGFR 측정값 존재함)

Category	Count	Predictors
Baseline	9	Age, Sugery position, Sex, Height, Weight, SBP, DBP, Hypertension, Diabetes
Lab	11	Serum uric acid, LDL, Triglycerid, Serum creatinine, eGFR, Cystatin-C, Cystatin-C eGFR, Creatinine clearance, Na (24hr urine), Volume (24hr urine)
CT	2	Remnant volume, Remnant volume percentage,
DTPA	4	Remnant predicted GFR, Normalized GFR, Remnant normalized GFR, Remnant relative uptake (%)
Derived variables	5	Bmi, Remnant Volume_bmi, Remnant Volume_height, Remnant Volume_weight, Remnant Volume_bsa

Regression models

Traditional regression models

- Mutiple Linear Regression
- Regression Splines

Machine learning

- Support Vector Machine
- Adaboost
- Random Forest
- Gradient boost
- Xtreme Gradient Boost

Traditional regression models

- 이론적으로 정규성, 독립성, 등분산성 등의 가정을 만족해야만 이러한 회귀 모델을 적용할 수 있음
- 하지만 대부분의 실제 임상 데이터는 이러한 엄격한 가정들을 만족하지 못함
- 임상데이터로 예측 모델을 적용하는 많은 연구들은 이러한 부분을 간과하고 있음

Machine learning

- 적용에 있어서 더 자유로운, 현대적인 머신러닝 알고리즘 기반의 예측모델을 함께 회귀분석에 적용함

1. 신장 기증환자 예후 예측

총 31개 변수로 시작 ➡ 변수 선별 ➡ 모델 트레이닝 & 모델 튜닝 ➡ 모델들의 성능지표 (MAE, RMSE) 합이 가장 낮은 변수선별 방식 선택 ➡ 해당 변수선별 방식에서 가장 좋은 성능의 모델 선택

feature selection

(1) 알고리즘 기반

- forward selection
- backward elimination
- stepwise selection
- lasso

(2) domain knowledge 기반

- 신장내과에서 선별해준 중요 변수

(3) 알고리즘 + domain knowledge

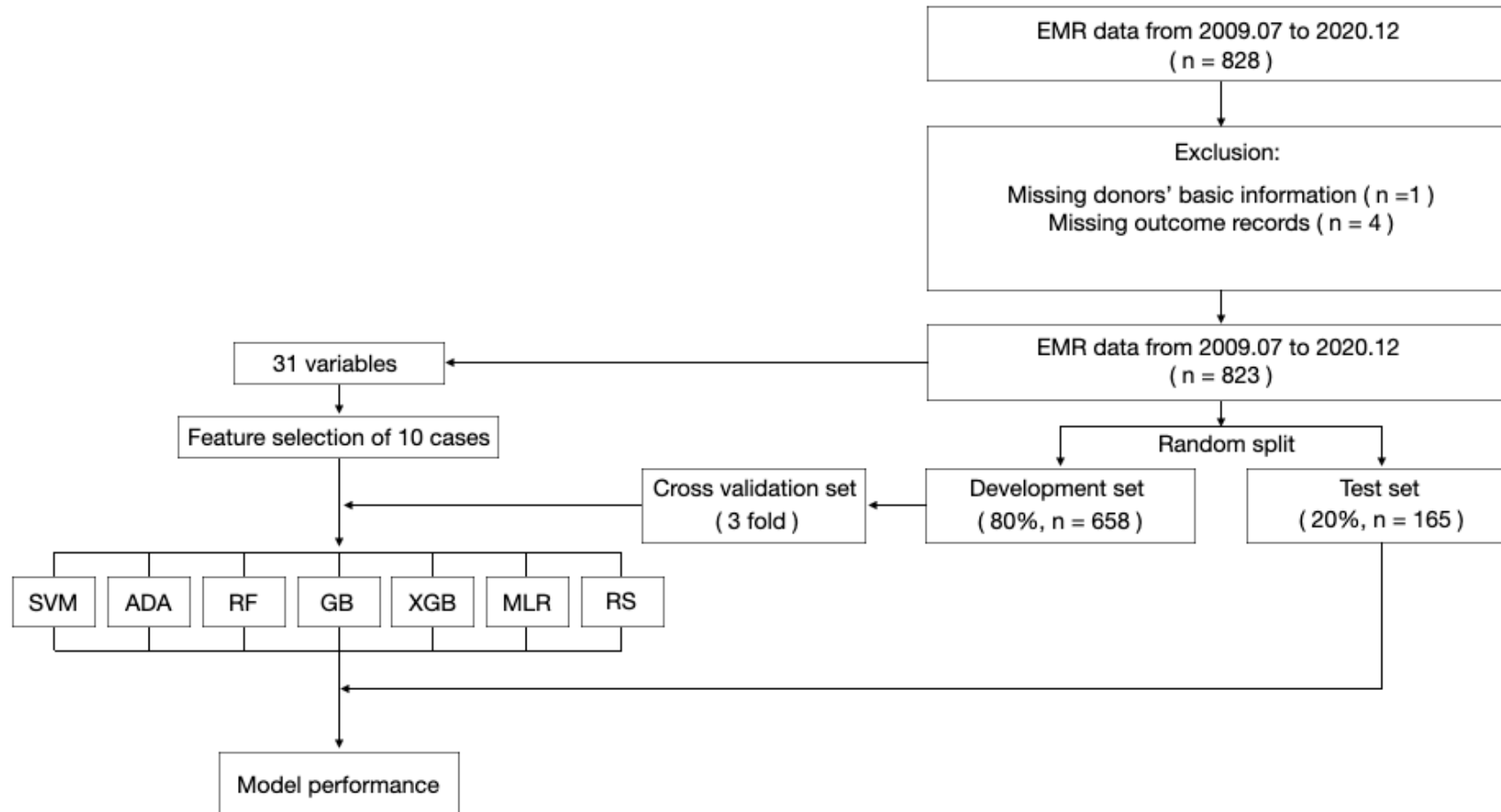
- forward selection + domain knowledge
- backward elimination + domain knowledge
- stepwise selection + domain knowledge
- lasso + domain knowledge

10 variable subsets

whole (31)
forward selection (12)
backward elimination (11)
stepwise selection (16)
lasso (10)
essential (13)
forward selection & essential (18)
backward elimination & essential (19)
stepwise selection & essential (18)
lasso & essential (21)

1. 신장 기증환자 예후 예측

< 실험 아키텍처 >



SVM = Support Vector Machine
ADA = Adaboost
RF = Random Forest
GB = Gradient boost
XGB = Xtreme Gradient Boost
MLR = Mutiple Linear Regression
RS = Regression Splines

1. 신장 기증환자 예후 예측

eTable 3. MAE (RMSE) performance of the predictive models

Models Methods	SVM	Random forest	Adaboost	GBoost	XGBoost	MLR [*]	RS [†]	Sum of MAE (RMSE)
Whole	7.50 (9.71)	6.48 (8.35)	6.76 (8.57)	6.82 (8.69)	6.46 (8.23)	6.47 (8.15)	6.48 (8.29)	46.97 (59.99)
Essential	6.67 (8.63)	6.71 (8.62)	6.81 (8.40)	6.63 (8.58)	6.49 (8.22)	6.49 (8.20)	6.45 (8.19)	46.16 (58.84)
Forward	6.30 (8.20)	6.29 (8.24)	6.54 (8.44)	6.63 (8.49)	6.23 (8.06)	6.33 (8.11)	6.64 (8.28)	44.96 (57.82)
Backward	6.41 (8.35)	6.35 (8.34)	6.67 (8.43)	6.45 (8.44)	6.59 (8.14)	6.29 (8.02)	6.50 (8.35)	45.00 (58.07)
Stepwise	8.30 (10.63)	6.45 (8.33)	7.03 (9.02)	6.60 (8.43)	9.07 (8.34)	6.34 (8.13)	6.51 (8.30)	47.82 (61.18)
Lasso	9.59 (11.73)	9.21 (11.23)	9.16 (11.19)	9.43 (11.52)	6.37 (10.98)	9.08 (11.12)	8.97 (10.99)	64.51 (78.76)
Essential & Forward	6.51 (8.38)	6.53 (8.40)	6.81 (8.61)	6.61 (8.46)	6.51 (8.16)	6.41 (8.10)	6.48 (8.31)	45.86 (58.42)
Essential & Backward	6.52 (8.42)	6.53 (8.39)	6.80 (8.57)	6.63 (8.46)	6.36 (8.16)	6.38 (8.05)	6.48 (8.31)	45.68 (58.36)
Essential & Stepwise	6.51 (8.38)	6.53 (8.40)	6.81 (8.61)	6.61 (8.46)	6.51 (8.16)	6.41 (8.10)	6.48 (8.31)	45.86 (58.42)
Essential & Lasso	6.53 (8.38)	6.69 (8.51)	7.01 (8.51)	6.58 (8.50)	6.55 (8.23)	6.55 (8.16)	6.43 (8.21)	46.30 (58.50)



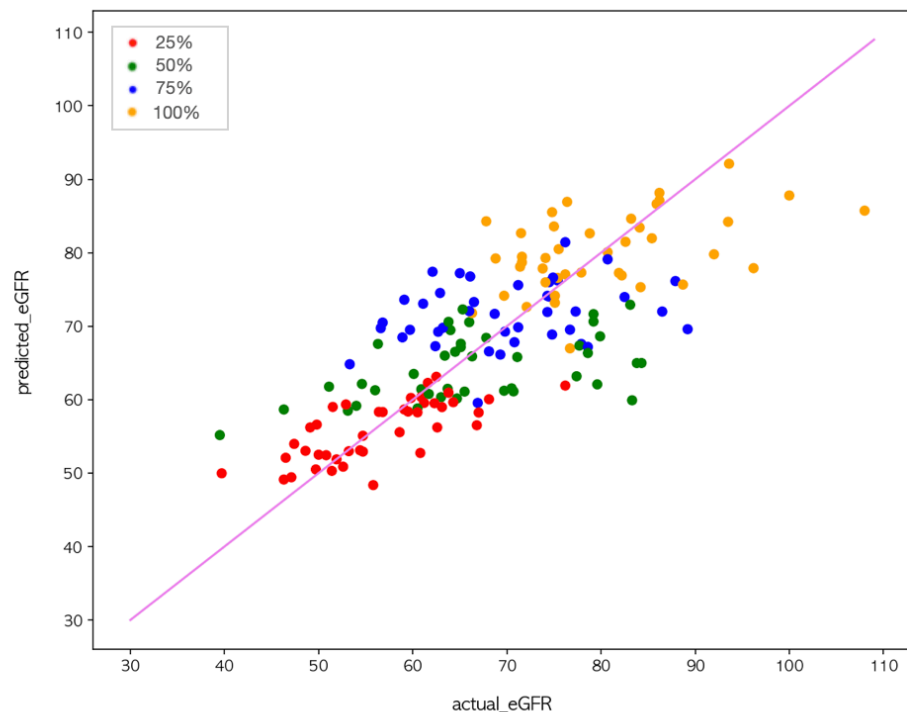
- 1) 가장 좋은 성능의 변수선별 방식: Forward selection
- 2) Forward 방식에서 가장 좋은 모델: XGBoost

^{*}MLR: Multi-variable Linear Regression

[†]RS: Regression Splines

1. 신장 기증환자 예후 예측

< Result - Scattor Plot >



Test set 으로 예측을 수행한 결과를 plot으로 나타낸 것

환자들의 기증 전 eGFR의 quantile을 색으로 표현함

빨강: 테스트 그룹에서 기증 전 eGFR이 하위 25% 이하인 사람들

초록: 테스트 그룹에서 기증 전 eGFR이 25%~50%인 사람들

파랑: 테스트 그룹에서 기증 전 eGFR이 50%~75%인 사람들

노랑: 테스트 그룹에서 기증 전 eGFR이 상위 75%이상인 사람들

< Result - Table (feature subset of Forward Selection) >

Table 4. Performance of predictive models

Metrics	MAE	RMSE	Predicted values within $\pm 5(\%)$ of the actual values	Predicted values within $\pm 10(\%)$ of the actual values	Predicted values within $\pm 15(\%)$ of the actual values	Predicted values within $\pm 20(\%)$ of the actual values
Models						
SVM	6.30	8.20	0.36	0.60	0.80	0.90
Random forest	6.29	8.24	0.41	0.59	0.78	0.90
Adaboost	6.54	8.44	0.37	0.59	0.75	0.87
GBoost	6.63	8.49	0.36	0.58	0.77	0.88
XGBoost	6.23	8.06	0.39	0.58	0.81	0.90
Multi-variable linear regression	6.33	8.11	0.33	0.64	0.81	0.90
Regression splines	6.64	8.28	0.36	0.60	0.79	0.90

- 테스트 데이터셋으로 테스트 한 결과
- Predicted values가 Actual values 와 비교했을때 평균적으로 몇 % 오차를 보이는지 제시함
- 임상적으로 몇 %까지 허용오차 범위인지 파악하면 더 정확한 성능 비교 가능.

1. 신장 기증환자 예후 예측

< Feature importance >

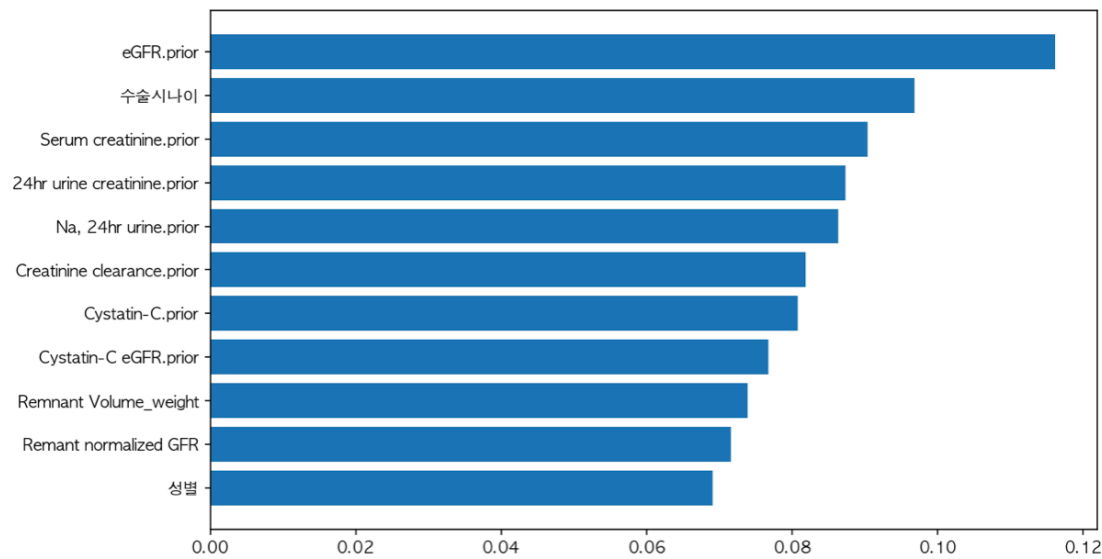


Figure 4. Feature importance of XGBoost model

< Donor's eGFR Calculator 스크린샷 >

KDNI (Kidney Donation with Nephrologic Intelligence) Smart Health Lab

Baseline information	Test result	CT result	DTPA result
수술시 나이 45	Serum uric acid 2.6	Lt.Kidney volume 112	Predicted GFR, Lt 58.36
성별 F	LDL 124	Rt.Kidney volume 134	Predicted GFR, Rt 51.3
Height (cm) 160	Triglycerid 120	Total volume 246	Predicted GFR, total 109.66
Weight (kg) 65.4	Serum creatinine 0.83		Normalized GFR 113.04
SBP 126	eGFR 85.4		상대섬위율(Lt, %) 53.22
DBP 88	Cystatin-C 0		상대섬위율(Rt, %) 46.78
Bmi 25.54	Cystatin-C eGFR 0		Remnant normalized GFR 52.9
	Creatinine clearance 79		
	24hr urine creatinine 0.9		
	Na, 24hr urine 101		
	Volume 24hr urine 1000		

submit

Figure 4. KDNI(Kidney Donation Nephrologic Intelligence) web applications

2. 다국가 기반 손상환자의 뇌 내 출혈 예측

1) 연구목표

- 국내 응급환자 데이터로 뇌 내 출혈(Intracranial hemorrhage, ICH) 예측모델 생성 (classification)
- 상대적으로 데이터가 부족한 다른 국가들에도 모델을 적용시키기 위한 전이학습 (Transfer learning) 수행

2) 기여도

- 현재 논문 게재 진행중 (1저자)
- 실험 결과에 대한 ppt 준비 및 발표
- 데이터 전처리, 머신러닝 기반의 데이터마이닝 분석 결과 제시

3) Skills

- python, scikit-learn, pandas, numpy, tensorflow (keras)

3) 분석과정 요약

a) EDA

- 변수 분류 (categorical, numeric)
- Basic demographic 확인

b) Data integration

- 국내 데이터와 아시아 데이터, 서로 다른 두 레지스트리의 변수 통합을 위한 명세서 작성
- python 환경에서 하나의 데이터프레임으로 통합

c) Data Preprocess

- exclusion criteria 적용: 사용 할 수 없다고 판단되는 데이터 제거 (환자 기본정보, 필수변수 누락된 경우 등..)
- 결측값 대체 (변수별로 오름차순/내림차순 정렬하면서 극단값 재변환, 또는 변수별로 일괄 대표값 대체)
- 모델링에 사용하기 위한 최종 데이터 생성 : demographic 최종 확인

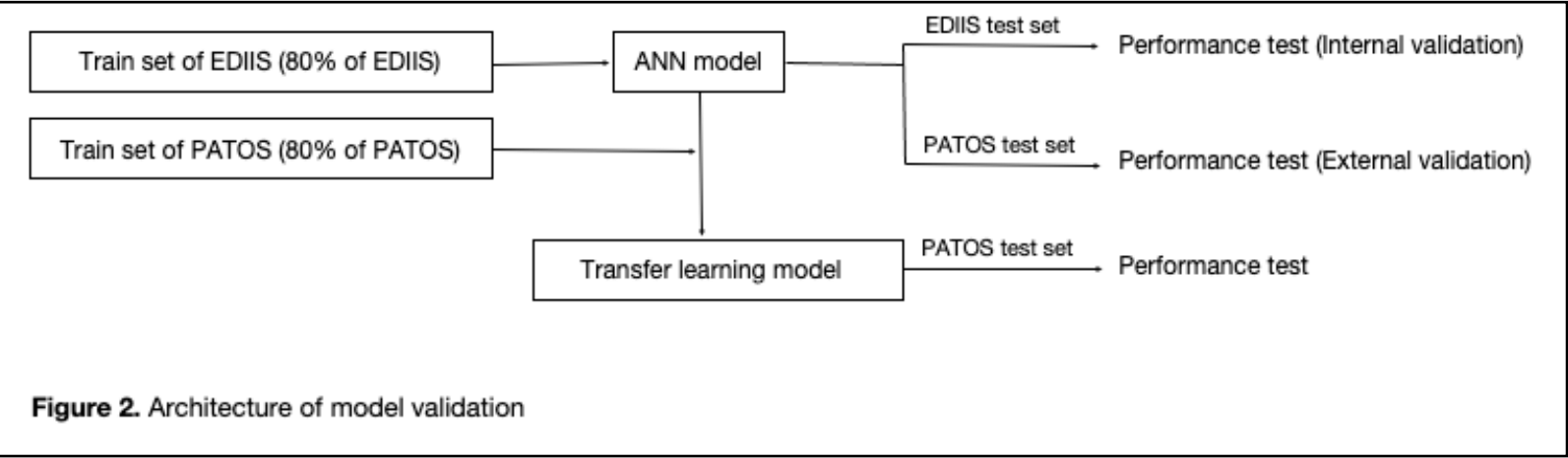
d) Modeling

- 머신러닝 모델 트레이닝
- 모델 튜닝: 모델별 하이퍼파라미터 튜닝

e) Result summary

- 성능지표 제시 (Table)

2. 다국가 기반 손상환자의 뇌 내 출혈 예측



국내 데이터(EDIIS)로 트레이닝한 ANN 모델의 성능평가.
(국내데이터(EDIIS)와 해외 다국가 데이터(PATOS)로 평가함)

Table 3. AUROC Score of Each Country.

	Internal validation	EDIIS model external validation in PATOS	EDIIS & PATOS Transfer learning ANN model
Korea (n = 11,058)	0.86 (0.85-0.87)	0.85 (0.84 – 0.87)	0.87 (0.85–0.88)
Malaysia (n = 4,343)	-	0.77 (0.73 – 0.81)	0.78 (0.74–0.82)
Vietnam (n = 4,133)	-	0.79 (0.73 – 0.85)	0.84 (0.80–0.88)
Japan (n = 368)	-	0.66 (0.42 – 0.85)	0.71 (0.52–0.88)
Taiwan (n = 1,152)	-	0.63 (0.38 – 0.90)	0.60 (0.26–0.90)

Values are described as AUROC scores (95% confidence interval).

Abbreviations: ANN: artificial neural network; AUROC: area under the receiver operating characteristic; EDIIS: Emergency Department-based Injury In-depth Surveillance; PATOS: Pan-Asian Trauma Outcome Study.

국가별로 ANN 모델의 성능을 평가함
Internal validation : EDIIS 데이터만으로 성능평가
EDIIS model external validation in PATOS : 국내 모델을 해외 데이터로 성능평가
EDIIS & PATOS Transfer learning ANN model : 전이학습 모델의 성능평가

3. 응급실 처치 추천모델의 효용성 평가

1) 연구목표

- 응급실에 도입되어 사용중인 처치 추천 모델의 도입 이전과 이후의 의료행위 효용성 변화 확인
- 병원 시스템에서 작동중인 예측 모델과 동일한 모델을 구현하고, 그 성능을 비교 평가함

2) 기여도

- 병원 시스템에서 작동중인 예측 모델과 동일한 모델을 구현하고, 그 성능을 비교 평가함
- 결과 테이블 도출

3) Skills

- python, scikit-learn, pandas, numpy

3) 분석과정 요약

- 프로젝트의 이전 기록들 확인
 - 병원에 도입한 모델에 대한 보고서 확인
 - 처치 추천에 사용된 threshold와 현재 적용한 모델 종류 확인
- Data Preprocess
 - 이전과 동일한 기준을 적용하여 데이터 전처리
- Modeling
 - 머신러닝 모델 트레이닝 및 튜닝
 - 최적 threshold 도출하고 보고서와 비교
- Result summary
 - 결과 리포트를 위한 Table 도출

3. 응급실 처치 추천모델의 효용성 평가

<Table 03>	min (IQR1, IQR3)	
	pre_Y	post_Y
Aline_Y		
time_Aline	95 (45, 271)	88 (47, 252)
time_O2	21 (7, 167)	22 (8, 111)
time_Highflow	324 (93, 655)	168 (69, 526)
time_Intubation	131 (71, 490)	132 (29, 402)
time_Inotropics	114 (48, 346)	176 (65, 416)
O2_Y		
time_Aline	92 (44, 271)	83 (46, 223)
time_O2	27 (8, 271)	28 (9, 181)
time_Highflow	332 (99, 685)	183 (70, 566)
time_Intubation	131 (74, 480)	143 (30, 428)
time_Inotropics	109 (44, 341)	167 (64, 413)
Highflow O2_Y		
time_Aline	84 (42, 236)	80 (45, 211)
time_O2	23 (7, 151)	25 (9, 140)
time_Highflow	332 (99, 685)	180 (70, 563)
time_Intubation	131 (74, 436)	149 (37, 435)
time_Inotropics	99 (43, 293)	144 (57, 382)
Intubation_Y		
time_Aline	85 (42, 237)	78 (44, 212)
time_O2	18 (7, 122)	21 (8, 89)
time_Highflow	322 (89, 663)	162 (68, 535)
time_Intubation	131 (71, 475)	143 (31, 432)
time_Inotropics	89 (42, 282)	144 (56, 383)
Inotropics_Y		
time_Aline	89 (44, 232)	85 (46, 228)
time_O2	21 (7, 155)	21 (8, 93)
time_Highflow	324 (97, 662)	152 (67, 517)
time_Intubation	131 (74, 456)	137 (30, 401)
time_Inotropics	117 (50, 349)	184 (67, 460)

<Table 04>	min (IQR1, IQR3)			
	pre_Y	post_Y	pre_N	post_N
Aline_(Y/N)				
time_Aline	95 (45, 271)	88 (47, 252)	242 (116, 467)	286 (111, 533)
time_O2	21 (7, 167)	22 (8, 111)	322 (103, 649)	370 (116, 919)
time_Highflow	324 (93, 655)	168 (69, 526)	670 (472, 904)	568 (142, 1338)
time_Intubation	131 (71, 490)	132 (29, 402)	236 (167, 833)	442 (156, 699)
time_Inotropics	114 (48, 346)	176 (65, 416)	433 (212, 683)	691 (325, 1166)
O2_(Y/N)				
time_Aline	92 (44, 271)	83 (46, 223)	249 (116, 498)	270 (113, 475)
time_O2	27 (8, 271)	28 (9, 181)	331 (110, 668)	367 (92, 936)
time_Highflow	332 (99, 685)	183 (70, 566)	714 (564, 904)	648 (473, 1338)
time_Intubation	131 (74, 480)	143 (30, 428)	503 (222, 942)	709 (462, 874)
time_Inotropics	109 (44, 341)	167 (64, 413)	436 (212, 691)	626 (328, 1080)
Highflow O2_(Y/N)				
time_Aline	84 (42, 236)	80 (45, 211)	247 (118, 490)	279 (111, 502)
time_O2	23 (7, 151)	25 (9, 140)	332 (112, 668)	370 (99, 946)
time_Highflow	332 (99, 685)	180 (70, 563)	733 (515, 924)	723 (496, 1491)
time_Intubation	131 (74, 436)	149 (37, 435)	372 (199, 924)	668 (126, 747)
time_Inotropics	99 (43, 293)	144 (57, 382)	422 (194, 683)	586 (270, 1059)
Intubation_(Y/N)				
time_Aline	85 (42, 237)	78 (44, 212)	249 (117, 498)	290 (115, 545)
time_O2	18 (7, 122)	21 (8, 89)	354 (122, 672)	386 (114, 972)
time_Highflow	322 (89, 663)	162 (68, 535)	670 (472, 902)	797 (488, 1643)
time_Intubation	131 (71, 475)	143 (31, 432)	372 (199, 924)	432 (195, 678)
time_Inotropics	89 (42, 282)	144 (56, 383)	484 (247, 693)	628 (292, 1062)
Inotropics_(Y/N)				
time_Aline	89 (44, 232)	85 (46, 228)	240 (116, 461)	286 (113, 505)
time_O2	21 (7, 155)	21 (8, 93)	329 (102, 658)	369 (97, 912)
time_Highflow	324 (97, 662)	152 (67, 517)	752 (529, 916)	569 (390, 1066)
time_Intubation	131 (74, 456)	137 (30, 401)	283 (210, 905)	215 (86, 566)
time_Inotropics	117 (50, 349)	184 (67, 460)	411 (167, 691)	586 (229, 1034)

Pre_Y : 모델 적용 이전, 처치를 수행 한 경우 소요된 시간
 Pre_N : 모델 적용 이전, 처치를 수행하지 않은 경우 소요된 시간
 Post_Y : 모델 적용 이후, 처치를 수행 한 경우 소요된 시간
 Post_N : 모델 적용 이후, 처치를 수행하지 않은 경우 소요된 시간

Aline
 O2
 Highflow
 Intubation
 Inotropics

→ 응급실에서 수행하는 몇 가지 처치들

4. CPR dashboard 개발

1) 연구목표

- 소생실의 CPR 현황판 (Dashboard) 데스크톱 어플리케이션 개발

2) 기여도

- dashboard 어플리케이션 개발
- android 앱과 네트워크 연동 (socket.io)

3) Skills

- node.js, javascript, react, electron



THANK YOU 😊