AUGMENTED INTELLIGENCE LAB

당뇨병 모델 제작

황승현

경상국립대학교 컴퓨터과학과
증강지능 연구실

01. WHAT I DID IN 2021

01

2022_NEW_SH

2022_NEW_승현 선언.
확 달라진 모습 보여드리겠습니다.
헤어스타일, 블로그, 1일 1백준
밀크씨슬, 헬스장

02

ESD HOTDEAL

2021-07-20부터 2021-11-24까지 개발한 ESD HotDeal이 2021년 12월 29일 정식으로 개발 종료
제1회 경남소프트웨어경진대회에 출품하여 최우수상(대학총장상) 수상
제30회 소프트웨어 전시회 입상

03

DIABETES

Classification and prediction on incidence of hypertension with blood pressure determinants in a deep learning model의 후속 연구
당뇨병 분석 모델 개발 시작

02. WHAT IS DIABETES?

DIABETES

당뇨병이란 혈액중의 포도당(혈당)이 높아서 소변으로 포도당이 넘쳐 나오는 것. 탄수화물은 위장에서 소화효소에 의해 포도당으로 변한 다음 혈액으로 흡수됩니다. 흡수된 포도당이 우리 몸의 세포들에서 이용되기 위해서는 인슐린이라는 호르몬이 반드시 필요합니다. 인슐린은 췌장 랑게르한스섬에서 분비되어 식사 후 올라간 혈당을 낮추는 기능을 합니다. 만약 여러 가지 이유로 인하여 인슐린이 모자라거나 성능이 떨어지게 되면, 체내에 흡수된 포도당은 이용되지 못하고 혈액 속에 쌓여 소변으로 넘쳐 나오게 되며, 이런 병적인 상태를 '당뇨병' 이라고 부르고 있습니다.

출처: 대한당뇨병학회

01 DRUGINSCU

인슐린 투여 여부

1. 아니요, 2예

하나라도 해당하면 당뇨병

02 GLUO_TR

공복 혈당

126mg/dl 이상

하나라도 해당하면 당뇨병

03 GLU120_TR

당부하 120분 후 혈당

200mg/dl 이상

하나라도 해당하면 당뇨병

03. DEVELOPMENT ENVIRONMENT

DEVELOPMENT ENVIRONMENT

Google Colab에서 개발

keras Sequential 모델

Python 3.7.12

tensorflow	2.7.0
numpy	1.19.5
pandas	1.1.5
matplotlib	3.2.2
google-colab	1.0.0
keras-tuner	1.1.0

04. DIFFERENCE FROM BEFORE

더 세밀한 데이터 전처리

이전 고혈압 분석 모델에서 부족했던 점을 찾아 개선하였고, 당뇨병 모델의 데이터에 맞게 적절하게 데이터 전처리를 한다.

fillna	pandas.DataFrame.fillna() 일부 변수의 결측값을 0으로 대체
변수 유형별 분류	binary / categoryH0 / categoryH1 / continuous
Group 분류	데이터를 성별, 나이에 따라 6가지로 그룹화 키, 몸무게 데이터의 결측값을 채울 때, 각 그룹의 평균으로 대치

TUNE THE TUNER

keras tuner 파라미터 조정
Dropout late, Learning late 고정
unit의 개수는 이전 고혈압 분석 모델을 참고

Sequential Model의 Hidden layer의 개수도 3으로 고정.
이전 고혈압 모델에서 가장 좋은 결과.

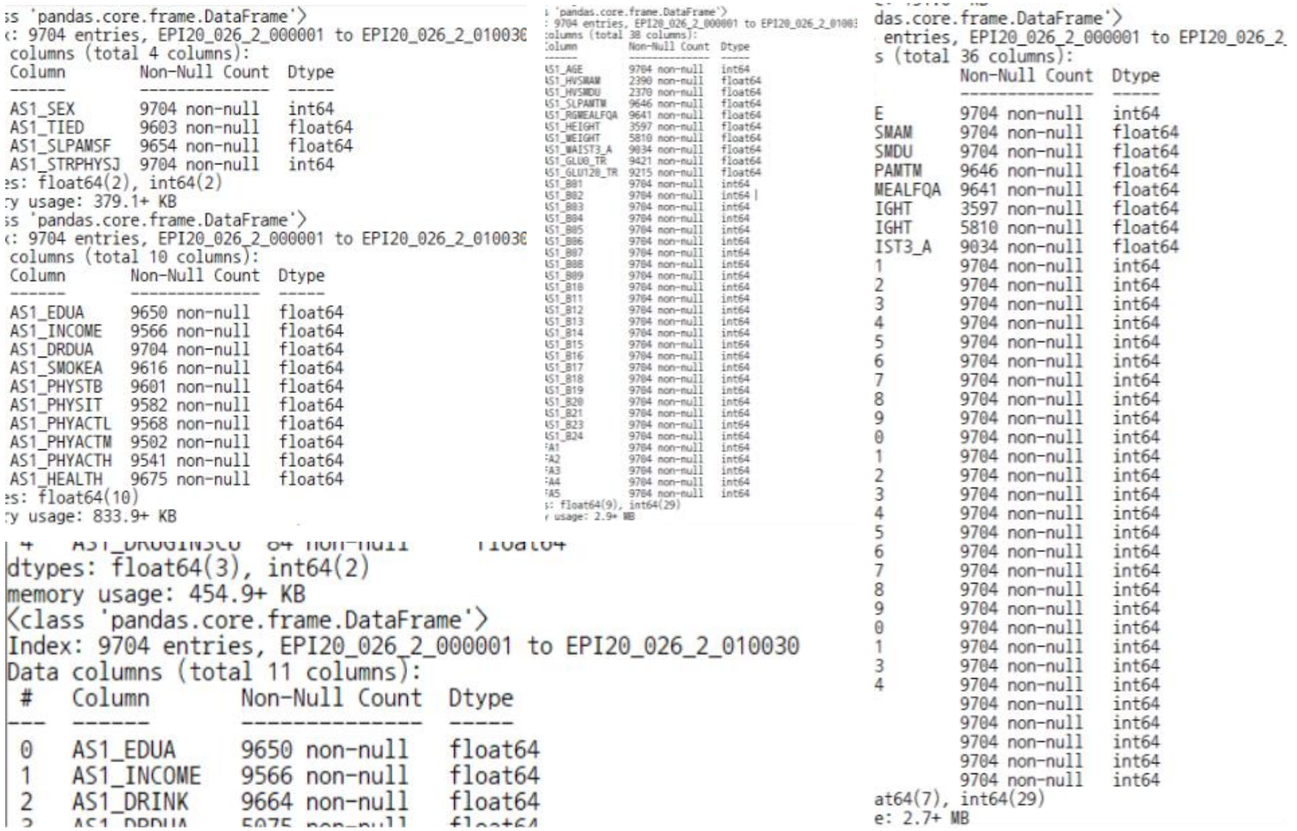
SAW DECISION TREE

Decision Tree / Random Forest 삭제
회귀계수로 변수 중요도 판단

회귀계수란?
회귀분석에서 독립변수가 한 단위 변화함에 따라 종속변수에 미치는 영향력 크기 두 변수 사이에 상관관계가 거의 없을 때 회귀계수는 의미가 없게 된다.

BEFORE & AFTER

당뇨병 모델에 맞게 변수의 수(column)을 줄이고, 변수의 결측값을 특성에 맞게 대치하였다. 예를 들어 AS1_HVSMAM, AS1_HVSMDU는 각각 습관적 흡연자의 하루 흡연량, 흡연 기간이다. 비흡연자는 해당 항목을 검사하지 않아서, 총 9704개 데이터에서 유효한 값(Non-Null Count)이 2370, 2390개 밖에 없었다. 이 값을 모두 0으로 대치하였다.



```
dtypes: float64(3), int64(2)
memory usage: 454.9+ KB
<class 'pandas.core.frame.DataFrame'>
Index: 9704 entries, EPI20_026_2_000001 to EPI20_026_2_010030
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   AS1_EDUA     9650 non-null   float64
 1   AS1_INCOME   9566 non-null   float64
 2   AS1_DRINK    9664 non-null   float64
```

05. MODEL DESIGN PLAN

01

독립변수 예측

기존과 같은 방법

DIABETES 독립변수 예측하는 모델 개발

02

모델 3개

AS1_DrugInsCu, AS1_Glu0_TR, AS1_Glu120_TR 을 예측하는 회귀 모델 1개 개발

03

다중 회귀 분석

AS1_DrugInsCu, AS1_Glu0_TR, AS1_Glu120_TR 을 각각 예측하는 선형 회귀 모델 개발

06. DEVELOPMENT STATUS

DEVELOPMENT STATUS

앞서 언급한 데이터 전처리, keras tuner의 파라미터 조정, Hidden l ayer의 수를 모두 조정하였다. 기본적인 구조는 기존의 고혈압 모델과 동일하다.

모델을 학습했을 때, 정확도는 0.8879까지 나왔다. 고혈압 학습모델이 78정도 나온 것을 생각하면 상당히 높은 정확도이다.

trial 1

> Trial 810 Complete [00h 00m 03s]
>
> val_accuracy: 0.884095311164856
>
> Best val_accuracy So Far: 0.8879587650299072
>
> Total elapsed time: 03h 00m 10s
>
> INFO:tensorflow:Oracle triggered exit
>
> unit: 12 learning_late: 0.001 dropout: 0.0.

trial 2

> Trial 29 Complete [00h 00m 02s]
>
> val_accuracy: 0.884095311164856
>
> Best val_accuracy So Far: 0.884095311164856
>
> Total elapsed time: 00h 01m 21s
>
> INFO:tensorflow:Oracle triggered exit
>
> unit: 40

07. NEXT WEEK'S PLAN

NEXT WEEK'S PLAN

데이터를 나이, 성별에 따라 그룹화

그룹화한 데이터의 키, 체중 결측값 대치

BMI 데이터 제작

모델 정확도 향상

강의 듣기, 정보처리기사, 토익 공부하기

THANK YOU

황승현

컴퓨터과학과 증강지능 연구실