고혈압 분석 모델

2021-09-10

컴퓨터과학과 황승현

목차

- 저번주에 한 것
  - 고혈압 분석 모델 소개
- 이번주에 한 것
  - Decison Tree 개념 설명
  - acc, val_acc의 차이 설명
  - 변경된 데이터셋으로 새로운 모델 제작
- 다음주에 할 것
  - 모델 하이퍼 튜닝
  - 랜덤 포레스트 공부
  - 식영 김혜림 박사님과 미팅

저번주에 한 것

개요

- 식품영양학과 김혜림 박사님
- 사람의 나이, 영양, 식이 등 생활 패턴과 특정 질환의 상관 관계 조사
  - 식이 패턴과 고혈압의 상관 관계를 집중 연구
- 고혈압 모델 제작
  - 새로운 변수(사람)의 고혈압 유병 여부 예측.
  - 현재는 정상혈압이지만, 이후 고혈압에 걸릴지 예측

모델 설명

- 고혈압 변수 가공
- 데이터 분리 및 전처리
- 결측값 대치
  - 기반 데이터 변수 수정
  - KNNImputer, SimpleImputer
- 변수 스케일링
  - StandardScaler, MinMaxScaler, QuantileTransformer
- 모델 제작 및 하이퍼파라미터 튜닝
  - Train, test 분리
  - Keras Tuner
  - Hyperband

모델 설명

- Decision tree 제작
  - DecisionTreeClassifier
  - Decision tree로 변수 중요도 추출
- 다양한 모델 설계
  - 변수 솎아내고 모델 설계
  - 식이패턴을 이용한 모델 설계
  - 결측값을 모두 대치한 모델 설계

이번주에 한 것

이번주에 한 것

Decision Tree란?

의사결정 규칙(논리합, 논리곱)을 트리 구조로 나타내어 전체 자료를 몇 개의 작은 집단(지금은 고혈압 발병 여부)으로 나누어서 분석하는 기법이다. 즉, 귀납추론을 한다. 결정이 가능한 특징(식이패턴)으로 답(고혈압)을 추론한다.

Decision Tree가 작은 집단으로 나누는 방법은 ID3와 CART 알고리즘이 있지만, 사이킷런에서는 CART를 사용한다. CART 알고리즘은 Gini 불순도 개념을 사용한다. 불순도란 말 그대로 얼마나 한 그룹(트리의 노드) 안에 얼마나 정보가 다양하게 있는지(고혈압에 걸린 사람과 그렇지 않은 사람이 다양하게 있는지)이다. CART 알고리즘은 원래 데이터를 둘로 나누었을 때, 불순도를 최대한 낮추는 속성과 해당 값을 찾아 Decision Tree를 만든다.

Feature Importance란?

Feature Importance가 0이면 Decision Tree를 만들 때 그 변수가 전혀 사용되지 않음을 의미하고, Feature Importance가 1이면 Decision Tree를 이 변수를 사용하면 완벽하게 예측할 수 있다는 뜻이다. 특성 중요도의 합은 1이다.
특성 중요도의 값이 낮다고 그 변수가 모델을 만들 때 유용하지 않다고 보기는 어렵지만, 특성 중요도의 값이 크면 변수가 모델을 만들 때 유용하다고 해석할 수 있다. 왜냐하면, Tree와 Sequential 회귀를 하는 알고리즘만 다를 뿐 한 가지의 결과로 수렴하기 때문이다.

accuracy	정확도	데이터의 실제값과 모델의 판정값을 비교하였을 때, 데이터가 실제 값과 모델이 판정한 값이 일치하는 경우(TN, TP)가 있고 일치하지 않는 경우(FP, FN)가 있다. 정확도는 모든 경우 중 실제 값과 모델이 판정한 값이 일치하는 경우만 나타낸 것이다.
acc	정확도	훈련 데이터(X_train)를 validation_split으로 나누었을 때 모델을 제작할 때 사용한 데이터로 평가한 정확도이다.
val_acc	검증(validation) 정확도	훈련 데이터(X_train)를 validation_split으로 나누었을 때 모델을 제작할 때 사용하지 않은(모델이 처음 보는) 데이터로 평가한 정확도
evaluate()	평가	모델의 내장 메소드(함수)이다. 모델의 loss(오차)와 accuracy(정확도)를 반환한다. evaluate 메소드에는 X_test와 Y_test (역시나 모델이 처음 보는 변수)를 넣어 모델을 판단한다. val_acc와 같다.

		모델의 판정값		
		음성	양성	
데이터의 실제 값	음성	진음성, TN	위양성, FP	정확도(accuracy) (TN+TP) / (TN+TP+FN+FP)
	양성	위음성, FN	진양성, TP	

- Decison Tree 개념 설명
  - Decision Tree의 원리
  - CART 알고리즘
  - 특성 중요도
- 정확도 판별법 설명
  - 정확도의 정확한 정의
  - acc와 val_acc의 차이
- 변경된 데이터셋으로 새로운 모델 제작

Decision Tree 정확한 개념

- **Decision Tree**의 정확한 개념
  - 김혜림 박사님이 요청한 것
- 의사결정 규칙을 트리 구조로 나타냄
  - 논리합, 논리곱
- 전체 자료를 몇 개의 작은 집단으로 나누어서 분석
  - 식이 패턴
  - 고혈압 여부
- 결정이 가능한 특징(식이패턴)으로 답(고혈압)을 추론한다.
  - 귀납추론

Decision Tree 알고리즘

- Decision Tree의 알고리즘
  - 김혜림 박사님이 요청한 것
- CART
  - 사이킷런에서 사용
  - Gini 불순도
  - 불순도를 최대한 낮추는 속성과 값으로 Tree 나눔
- 불순도란?
  - 얼마나 정보가 다양하게 있는지

```
[21]  1 from sklearn.tree import DecisionTreeClassifier
      2 from sklearn.tree import export_graphviz
      3 import graphviz
      4
      5 ht_tree = DecisionTreeClassifier(max_depth=5)
      6 ht_tree.fit(X_test, y_test)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=5, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

```
[22]  1 export_graphviz(ht_tree, out_file="tree.dot",
      2                 class_names='FT',
      3                 feature_names=X.columns,
      4                 impurity=True, filled=True)
```

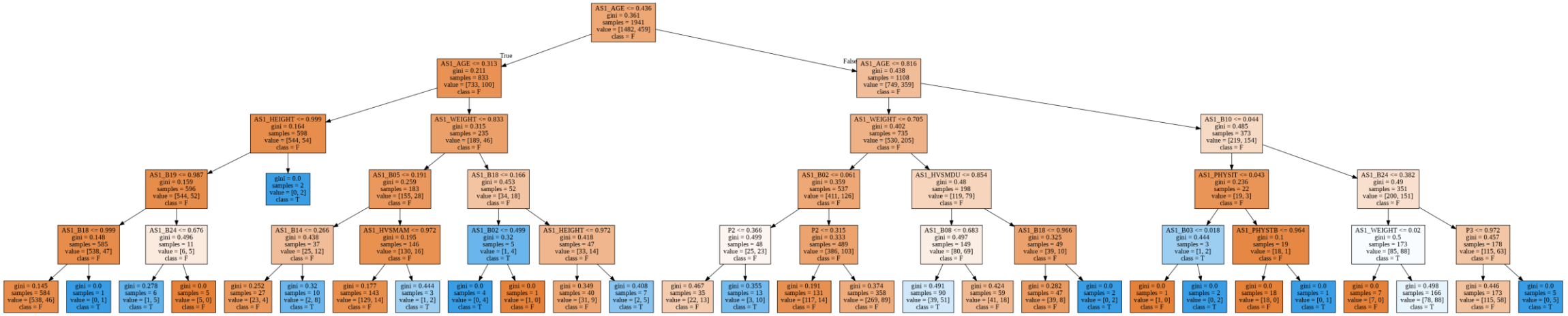
```
[17]  1 !ls
```

```
drive  sample_data  tree.dot
```

```
[23]  1 with open("tree.dot") as f:
      2     dot_graph = f.read()
      3
      4 graphviz.Source(dot_graph, filename='tree.png',format='png')
```

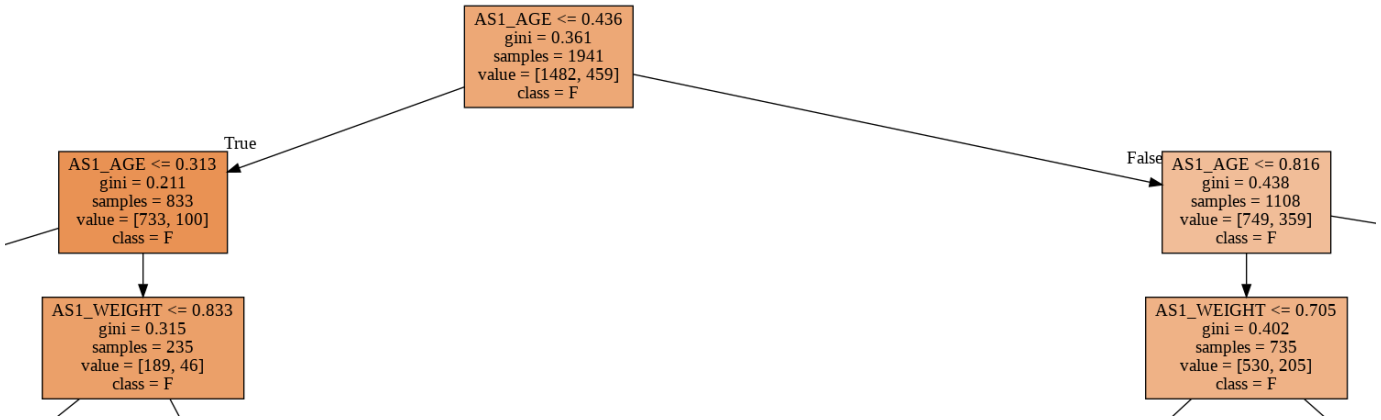
Decision Tree

- fit()
  - 의사결정트리 생성
- Export_graphviz()
  - 트리를 .dot 파일로 내보냄
  - 트리를 .png 파일로 내보냄



Decision Tree (max_depth = 5)

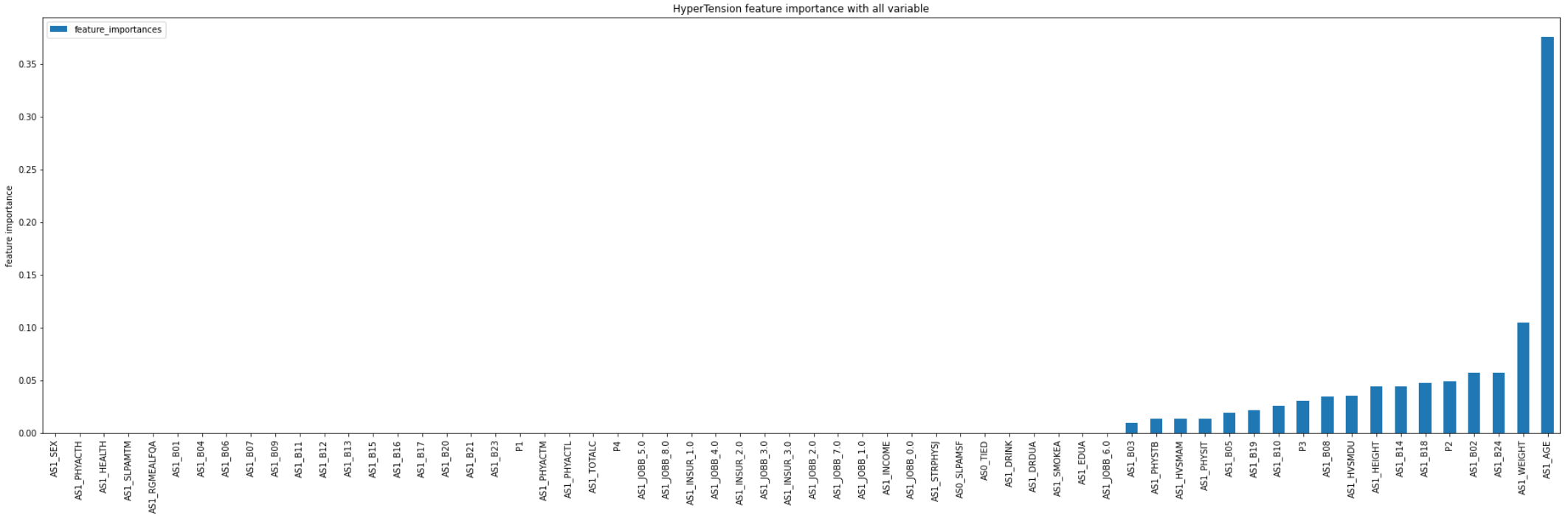
Decision Tree의 해석



- AS1_AGE <= 0.436
  - 분할조건
  - 노드를 쪼개는 기준
- Gini = 0.361
  - 노드의 지니 불순도
- Samples = 361
  - 노드에 있는 정보의 개수
- Value = [1482, 459]
  - Decision Tree로 분리한 샘플의 수

Feature Importance

- 특성 중요도
- Decision Tree를 만들 때 변수를 사용한 정도
  - 0: Decision Tree를 만들 때 그 변수가 전혀 사용되지 않음
  - 1: Decision Tree를 이 변수를 사용하면 완벽하게 예측할 수 있음
- 모든 변수의 Feature Importance의 합은 1
- 특성 중요도의 값이 낮으면?
  - 그 변수가 모델을 만들 때 유용하지 않다고 해석할 수는 없다.
  - 다른 변수만 가지고도 Decision Tree를 만들 수 있기 때문.



Feature importance

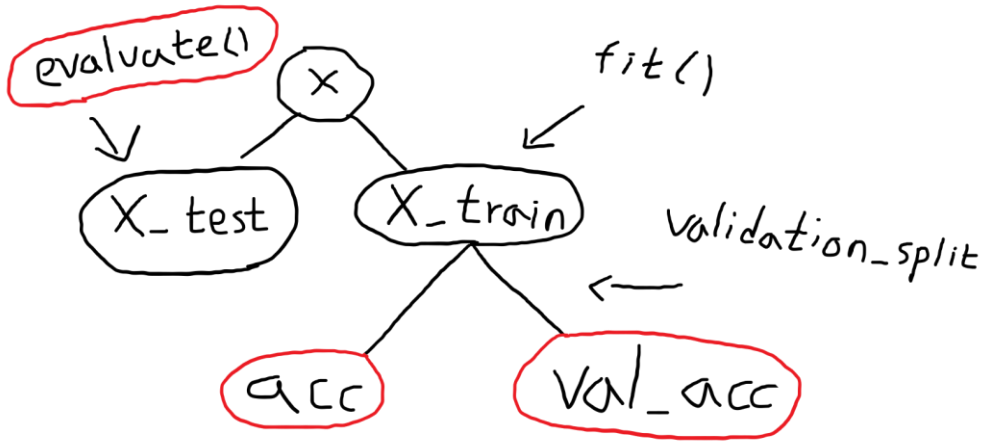
- max_depth = 5

accuracy

		모델의 판정값		정확도(accuracy) (TN+TP) / (TN+TP+FN+FP)
		음성	양성	
데이터의 실제 값	음성	진음성, TN	위양성, FP	
	양성	위음성, FN	진양성, TP	

- 정확도
- 데이터의 실제 값과 모델이 판정한 값이 일치하는 경우
  - 진음성, 진양성

정확도 판별법



- acc
  - 모델을 제작할 때 사용한 데이터로 평가한 정확도
- val_acc
  - 모델을 제작할 때 사용하지 않은 데이터로 평가한 정확도
- evaluate()
  - 하이퍼 모델의 내장 메소드. 모델의 loss, accuracy 반환
  - X_test, y_test를 넣어 모델 판단

다음주에 할 것

하이퍼 튜닝

- 데이터셋 변경
  - FA1 ~ FA5
- Hidden Layer 수 줄이기
  - 현재는 5개
- Epoch 수 늘리기
  - 현재는 최대 32
- Optimizer 바꾸기
  - 현재는 Adam
- 랜덤 포레스트 써보기

랜덤 포레스트

- 많은 Decision Tree로 분류 또는 평균 예측치 출력
- 다수의 Decision Tree 학습
  - ensemble
- bagging
  - 임의로 트리를 추출하여 생성, 중복허용
  - 전체 속성의 일부만 추출(Bagging Features)
- 장점
  - 월등히 높은 정확성
  - overfitting 확률 줄임, 일반화된 트리 생성
- sklearn.ensemble, RandomForestClassifier

끝없는 미팅

- 매주 주말
- 이번주에 보낸 자료 안내
- 다음주에 보낼 자료 전달
- 랜덤 포레스트 도입 제안

감사합니다~