퍼블릭 클라우드 서비스를 활용한 파이썬 기반 AL-ML

데이터 분석 프로젝트 보고서

날짜	2023. 05. 09
팀명	소수정예
이름	구영인

프로젝트 주제 :
수면 장애(Sleep Disorder)에 영향을 끼치는 요인 분석

1. 수면 장애를 일으키는 특성을 모은 데이터를 활용해 어떤 특성이 가장 많은 수면 장애에 영향을 끼치는지 분석하고자 함.
2. 주어진 데이터에서 전처리 후 표준화를 진행하였음.
3. 특성을 가장 잘 찾아내는 모델을 찾기 위해 3개의 모델을 사용하여 학습, 테스트를 진행함

▶ 데이터 전처리

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Person ID                374 non-null    int64
 1   Gender                   374 non-null    object
 2   Age                      374 non-null    int64
 3   Occupation               374 non-null    object
 4   Sleep Duration           374 non-null    float64
 5   Quality of Sleep         374 non-null    int64
 6   Physical Activity Level  374 non-null    int64
 7   Stress Level             374 non-null    int64
 8   BMI Category             374 non-null    object
 9   Blood Pressure           374 non-null    object
 10  Heart Rate               374 non-null    int64
 11  Daily Steps              374 non-null    int64
 12  Sleep Disorder           374 non-null    object
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
```

- 전체 데이터 확인 후 [Person ID]는 인덱스로 처리
- 회귀모델에서 사용하기위해 object 데이터들 중 [Gender, BMI Category]를 수치화 하기 위해서 원핫인코딩 방식을 사용하여 0, 1로 표현
- 혈압을 나타내는 [Blood Pressure]는 슬래시(/)를 기준으로 수축기와 이완기로 문자열 처리가 되어있어 실제 혈압에서는 수축기가 혈압 측정에 더 유효한 수치로 판단, 슬래시 이하 이완기는 삭제하여 수축기 데이터만 남도록 함
- 요인에 크게 작용하기 힘들 것이라고 판단한 [Occupation]과 변환된 데이터의 기존 데이터들은 데이터 셋에서 삭제
- [Sleep Disorder]의 경우 회귀모델에서 target 데이터로 사용하기 위해 선형회귀 모델을 사용 할 때만 원핫인코딩을 사용하여 0, 1로 표현
- 전처리한 데이터는 각 학습 데이터와 타겟 데이터로 나눠서 저장 후 train_test_split 함수를 이용하여 학습 데이터와 테스트 데이터로 구분
- 각 특성들의 수치가 차이가 있다고 판단하여 StandardScaler를 이용해 표준화 작업 진행

1. 선형 회귀 모델

```
from sklearn.linear_model import LinearRegression

# 표준화한 데이터 학습
lr = LinearRegression()
lr.fit(train_scaled, train_target)

print(lr.score(train_scaled, train_target))
print(lr.score(test_scaled, test_target))
```

```
0.5915987412437881
0.5856466165662247
```

```
# 표준화 하지 않은 데이터 학습
lr2 = LinearRegression()
lr2.fit(train_input, train_target)

print(lr2.score(train_input, train_target))
print(lr2.score(test_input, test_target))
```

```
0.5930226692367 82
0.5879419739363 25
```

▶ 분석 전 예상

- 수면 장애를 일으키는 요인이 종합되어 수면 장애 판단에 영향을 미칠 것이라고 예상했음
- 데이터에서 제공하는 수면 장애유형은 총 3가지로 예측모델(회귀모델)을 사용하여 각 특성의 수치를 입력하면 예측이 될 것 이라고 생각함

▶ 선형 회귀 모델 분석 후

- 그러나 실제로 학습 후 테스트 점수 확인 결과 선형 회귀 모델에서 예상과는 다르게 점수가 많이 낮게 나옴
- 표준화 한 lr모델과 표준화 하지 않은 lr2의 모델의 스코어가 거의 같음을 확인 할 수 있음
- 이 결과로 회귀모델에서는 분석이 제대로 되지 않는다는 것을 알 수 있었음
- 따라서 회귀모델이 아닌 분류모델에서 모델 학습과 테스트를 진행하기로 결정함
- 해당 과정에서 수면 장애 예측이 아닌 수면 장애 유형을 찾아내는 특성의 중요도를 분석하여 어떤 특성이 수면 장애에 가장 영향을 미치는지 분석하고자 함

2. 로지스틱 회귀 모델

```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()
lr.fit(train_scaled, train_target)

print(lr.score(train_scaled, train_target))
print(lr.score(test_scaled, test_target))
```

```
0.9178571428571428
0.9042553191489362
```

```
# 규제와 반복회수 지정
lr2 = LogisticRegression(C=20, max_iter=1000)
lr2.fit(train_scaled, train_target)

print(lr2.score(train_scaled, train_target))
print(lr2.score(test_scaled, test_target))
```

```
0.9214285714285714
0.9148936170212766
```

▶ 분석 전 예상

- 회귀모델보다 분류 모델을 이용해 분석해보는게 좋을 것 같다고 예상함
- test_size를 늘리면 더 좋은 모델 점수가 나올 것이라고 예상했음

▶ 로지스틱 회귀 모델 분석 후

- 오히려 테스트 사이즈가 늘어날수록 점수가 낮아지는 경향을 보였음
- 다섯 모델 중 규제를 적용했던 모델(lr2)에서 좋은 점수가 나온 것을 확인 할 수 있었음
- 해당 모델에서는 규제를 적용한 모델이 가장 좋은 점수를 내고 있다고 할 수 있음
- 분류 모델 예측 결과(predict) 5개 중 4개를 맞추는 모습으로 준수하다고 볼 수 있음

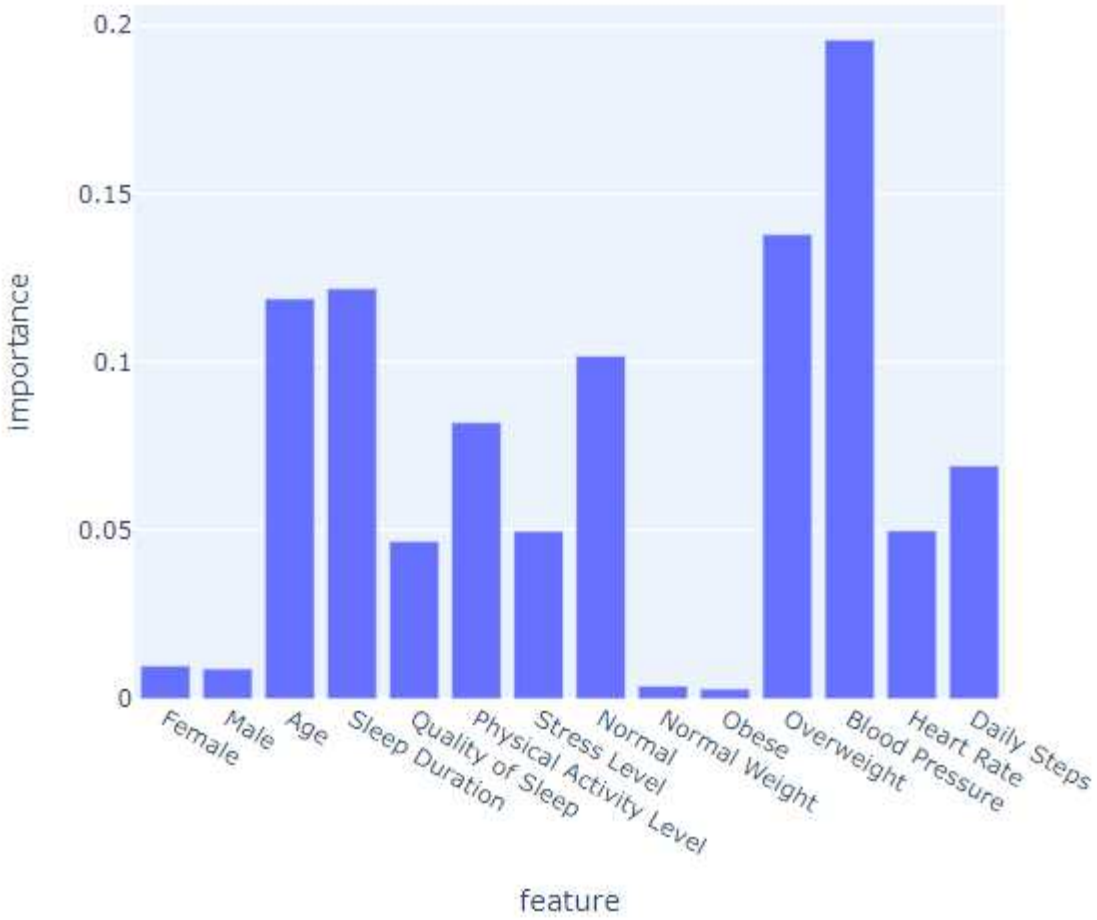
3. 랜덤 포레스트 모델

```
from sklearn.model_selection import cross_validate
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_jobs=-1, random_state=42)
scores = cross_validate(rf, train_input, train_target,
                        return_train_score=True, n_jobs=-1)

np.mean(scores['train_score']), np.mean(scores['test_score'])
```

```
(0.9339285714285713, 0.9035714285714287)
```

```
rf2 = RandomForestClassifier(n_jobs=-1, random_state=42)
scores2 = cross_validate(rf2, train_input, train_target,
                         return_train_score=True, n_jobs=-1)

np.mean(scores['train_score']), np.mean(scores['test_score'])
```

```
(0.9339285714285713, 0.9035714285714287)
```



특성 중요도 그래프

▶ 분석 전 예상

- 랜덤 포레스트 모델에 가장 준수한 점수를 낼 것이라고 예상함
- 랜덤 포레스트 특성 중요도를 이용해서 데이터의 특성 중 큰 요인을 뽑을 수 있을 것이라고 예상함

▶ 랜덤 포레스트 모델 분석 후

- rf 모델 : 테스트 사이즈를 지정하지 않은 기본 모델
- rf2 모델 : 테스트 사이즈를 0.3으로 지정한 모델
- 두 모델은 테스트 사이즈를 각각 다르게 지정하였지만 교차검증 시 결과 점수가 거의 비슷하다는 것을 확인함
- 테스트 사이즈 조정으로는 각 모델의 점수가 달라지지 않는 것을 알 수 있음
- 따라서 더 의미있는 점수를 얻기 위해서는 트리의 개수나 학습률을 조정해보는 등의 방법이 있을 것이라고 생각함
- 특성 중요도는 ['Blood Pressure', 'BMI Catagory', Sleep Duration'] 순으로 높음

▶ 분석 결과

- 결과적으로 회귀모델이 아닌 분류모델을 사용하여 학습했을 때 좋은 점수와 결과를 볼 수 있었음
- 선형 회귀 모델은 학습 데이터 점수와 테스트 점수의 차이가 별로 나지 않았지만, 자체적으로 점수가 너무 낮아서 평가의 지표로 삼기에는 어렵다고 판단했음
- 로지스틱 회귀 모델에서는 여러번의 하이퍼 파라미터 튜닝을 진행하였는데, 생각보다 테스트 사이즈의 크기와는 관계가 없었다는 것을 알 수 있었음
- 테스트 사이즈 크기보단 규제와 반복횟수를 지정한 모델이 가장 좋은 점수를 나타냈음
- 랜덤 포레스트 모델에서는 특성 중요도를 뽑아내고, 교차 검증을 통해서 해당 모델에서도 여러번 반복해서 학습과 검증을 거쳤을 때 좋은 점수가 나오는 것을 확인 할 수 있었음
- 가장 점수가 좋았던 모델은 규제와 반복횟수를 지정한 로지스틱 회귀 모델로 이러한 결과를 바탕으로 로지스틱 회귀 모델 뿐만 아니라 랜덤 포레스트나 다른 분류 모델에서도 반복횟수 지정, 트리 모델의 경우 트리의 개수 증가, 학습률 조정 등으로 더 좋은 점수를 얻을 수 있을 것이라고 예상함
- 이 과정을 통해 BMI 지수, 혈압, 수면시간 이 세 요인이 수면 장애 가장 높은 영향을 끼친다고 예상할 수 있음