[기계학습 및 실습] 기말시험 대체 과제

* 과제 작성 방법

1. 제출 파일은 2개입니다. : 보고서 파일(pdf), 프로그램 파일(텍스트 파일)

2. 보고서 : pdf 파일로 제출, 학번_이름.pdf 로 작성해서 제출
   - 실습문제
     (1) 아래 문제에 첨부된 내용들이 보고서에 포함되어야 합니다.
     (2) 실습문제에 대하여 최적의 결과를 찾기 위하여 다양한 분석을 시도하여 추가한 advantage가 있습니다.
     (3) 실습문제에 대해서는 '파이썬프로그램'과 '실행결과'가 포함되어야 합니다. '파이썬프로그램' 또는 '실행결과'만 있으면 0점 처리합니다.

   - 그 외 문제들은 답안을 문항별로 작성합니다.

3. '파이썬프로그램' 은 텍스트 파일로도 제출해주세요.
   - 작성한 프로그램 파일을 메모장에 붙여넣어 저장하면 됩니다.
   - 프로그램 파일 이름 : 학번_이름.txt 또는 학번_이름.py로 작성해서 제출해주세요.

4. 제출기한 : 2023년 6월 22일(목) 밤11시 59분까지
   - 지각제출시 0점 처리합니다.

1.(20점)"hospital.txt" 데이터는 미국 내 113개의 병원들을 대상으로 입원기간 동안 환자들이 받는 감염위험과 관련된 사항들을 조사하였다. 다음은 hospital 데이터에 관한 정보이다.

```
hospital.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113 entries, 0 to 112
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   ID          113 non-null    int64
 1   Stay        113 non-null    float64
 2   Age         113 non-null    float64
 3   InfctRsk    113 non-null    float64
 4   Culture     113 non-null    float64
 5   Xray        113 non-null    float64
 6   Beds        113 non-null    int64
 7   MedSchool   113 non-null    int64
 8   Region      113 non-null    int64
 9   Census      113 non-null    int64
 10  Nurses      113 non-null    int64
 11  Facilities  113 non-null    float64
dtypes: float64(6), int64(6)
memory usage: 10.7 KB
```

(1) 감염위험변수(InfctRsk)를 종속변수로 하여 독립변수들간의 산점도 행렬을 그리시오.

(2) 감염위험변수(InfctRsk)와 독립변수들간의 상관계수 행렬을 구하시오.

(3) 감염위험변수(InfctRsk)를 종속변수로 하고 나머지 11개 변수들을 독립변수로 하여 머신러닝 기반의 선형회귀분석을 수행하시오.

① 결측치 여부 확인
② 데이터 전처리 필요 여부 확인
③ 훈련용, 테스트용 데이터셋 분리
④ 선형회귀분석 결과를 토대로 감염위험에 대한 회귀식 작성: 절편과 회귀계수 구하기
⑤ MAE, MSE , RMSE, R2 평가지표를 통해 선형회귀분석 모델 평가
⑥ 회귀분석 결과를 산점도와 선형회귀 그래프로 시각화
⑦ 분석내용에 대하여 설명

(4) (3)에서 구한 11개의 독립변수들을 이용한 최종 회귀식에서 회귀계수가 0.01 이하인 변수를 제외하여 감염위험변수(InfctRsk)를 종속변수로 하고 나머지 변수들을 독립변수로 하여 머신러닝 기반의 선형회귀분석을 수행하시오.
① 감염위험에 대한 회귀식 작성: 절편과 회귀계수 구하기
② MAE, MSE , RMSE, R2 평가지표를 통해 선형회귀분석 모델 평가
③ 회귀분석 결과를 산점도와 선형회귀 그래프로 시각화
④ 분석 내용에 대한 설명

(5) (3) 과 (4) 의 결과를 비교하고 설명하시오.

2.(20점) "handspan.txt" 파일에서 키(Height)와 손 한뼘의 길이(HandSpan) 변수를 이용하여 남성인지 여성인지 분류하는 로지스틱 회귀를 수행하시오.
① 결측치 여부 확인
② 데이터 전처리 필요 여부 확인
③ 훈련용, 테스트용 데이터셋 분리
④ 로지스틱 회귀분석 결과를 토대로 모형식의 절편과 회귀계수 구하기
⑤ 평가데이터를 이용하여 예측값 구하기
⑥ 파이썬에서 오차행렬을 구하고 오차행렬을 이용하여 정확도, 정밀도, 재현율, F1 스코어, ROC 기반 AUC 스코어를 직접 계산하기
(※ 참고 : 데이터에서 성별 자료가 문자형 값으로 주어져 있어 파이썬에서 정확도, 정밀도, 재현율, F1 스코어 구할때는 `precision_score(Y_test, Y_predict, pos_label='Female')` 와 같이 `pos_label='Female'`을 추가해서 구할수 있으나 roc_auc score는 문자형을 숫자형으로 변환해야 구할 수 있음)

⑦ 분석내용에 대하여 설명

3.(30점) seaborn라이브러리에서 'mpg' 데이터를 로드하여 mpg 데이터프레임으로 저장하자. 다음은 mpg데이터프레임에 대한 정보이다.

```
mpg.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 398 entries, 0 to 397
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   mpg           398 non-null    float64
 1   cylinders     398 non-null    int64
 2   displacement  398 non-null    float64
 3   horsepower    392 non-null    float64
 4   weight        398 non-null    int64
 5   acceleration  398 non-null    float64
 6   model_year    398 non-null    int64
 7   origin        398 non-null    object
 8   name          398 non-null    object
dtypes: float64(4), int64(3), object(2)
memory usage: 28.1+ KB
```

'origin 변수는 자동차 제조국으로 usa, europe, japan 세 가지 값을 갖는다.
mpg, cylinders, displacement, horsepower, weight, acceleration, model_year 변수를 이용하여 자동차 제조국을 분류하는 결정트리 분석을 수행하시오.

① 결측치 여부 확인
② 데이터 전처리 필요 여부 확인
③ 훈련용, 테스트용 데이터셋 분리
④ 결정트리 분류분석 모델 구축하여 모델을 생성하고 트리 모형 적합(훈련)하고 예측 수행
⑤ 생성된 결정 트리 모델의 분류 정확도 성능을 확인
⑥ GridSearchCV모듈을 사용하여 정확도를 검사하고 최적의 하이퍼 매개변수를 찾는 작업 수행
⑦ Decision Tree Classifier의 주요매개변수들을 이용하여 조정하면서 최고의 평균 정확도 찾기
⑧ 최적 모델 grid_cv.best_estimator_을 사용하여 테스트 데이터에 대한 예측을 수행
⑨ feature_importances_ 속성을 사용하여 각 피처의 중요도를 알아내고 중요도가 높은 5개 피처를 찾아 그래프로 표시
⑩ Graphviz 패키지 : 결정트리 모델의 트리구조를 그림으로 시각화하기
⑪ 분석 내용에 대한 설명

4.(10점) 다음과 같은 수식을 사용하여 데이터 스케일링을 한다면 A열, B열에 있는 각 값들은 어떻게 변환되는지 작성하시오.

	A	B	C
0	14.00	103.02	big
1	90.20	107.26	small
2	90.95	110.35	big
3	96.27	114.23	small
4	91.21	114.68	small

$$z_i = \frac{x_i - \min(x_i)}{\max(x) - \min(x)}$$

5.(5점)) 스케일링 방법 중 z-스코어 정규화(z-score normalization)와 최솟값-최대값 정규화(min-max normalization)의 차이점에 대해서 설명하시오.

6.(5점) 다음과 같은 데이터셋을 파이썬 데이터프레임으로 작성하고 ①~④ 각 코드를 실행 결과를 적으시오.

데이터셋

	source	target	weight	color
0	0	2	3	red
1	1	2	4	blue
2	2	3	5	blue

① pd.get_dummies(edges).iloc[:,3:]
② pd.get_dummies(edges["color"])
③ pd.get_dummies(edges[["color"]])
④ pd.get_dummies(edges["color"], prefix="color")

7.(3점) 다음 코드를 실행할 경우 결과값을 적으시오.

```
import numpy as np
import pandas as pd

raw_data = {'first_name': ['Jason', np.nan, 'Tina', 'Jake', 'Amy'],
        'last_name': ['Miller', np.nan, 'Ali', 'Milner', 'Cooze'],
        'age': [42, np.nan, 36, 24, 73],
        'sex': ['m', np.nan, 'f', 'm', 'f'],
        'preTestScore': [4, np.nan, np.nan, 2, 3],
        'postTestScore': [25, np.nan, np.nan, 62, 70]}
df = pd.DataFrame(raw_data, columns = [
                                    'first_name', 'last_name', 'age', 'sex',
                                    'preTestScore', 'postTestScore'])

df.isnull().sum()
```

8.(7점) 다음 데이터에 대한 예측피쳐는 PLAY GOLF이다. 이 데이터에서 예측피쳐에 대한 엔트로피를 구하시오.

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No