< 단답형 및 서술형 >

1) 데이터 예측을 위한 머신러닝 모델의 구축 시 결측 데이터는 모델의 학습을 방해하는 요소이다. 결측 데이터를 처리하기 위한 전처리 방안을 4개 이상 작성하시오.

1. 제거 (컬럼 자체를 삭제, 레코드 단위로 제거)
2. 대치 (기초 통계 정보를 사용, 평균, 최빈값, 중앙값)
3. 머신러닝을 적용한 예측값으로 대치 (지도학습 기반)
4. 머신러닝을 적용한 예측값으로 대치 (비지도, 준지도학습 기반)

2) 선형회귀 분석을 위한 LinearRegression 클래스의 장점과 단점을 서술하시오. (2줄 이내)

- 장점 : 데이터에 과적합 되어 높은 학습 성능을 가질 수 있으므로 제약 조건이 추가된 Ridge, Lasso 클래스의 결과에 대해서 기준값이 될 수 있음
- 단점 : 데이터에 과적합되는 경향을 보이므로 특정 특성에 대해서 민감한 모델이 생성될 수 있음 (가중치가 높은 컬럼에 대해서)

3) 분류 분석 모델의 평가 지표 중 정밀도와 재현율에 대해서 설명하고, 다음의 CONFUSION MATRIX에서 정밀도와 재현율의 값을 클래스 1에 대해서 계산하시오.

	예측 0	예측 1
실제 0	75	9
실제 1	13	85

- 정밀도 : 머신러닝 모델이 예측한 값의 정답 비율
- 재현율 : 실제 데이터에서 머신러닝 모델이 정답으로 예측한 값의 비율
- 정밀도 값 : 85 / (9 + 85)
- 재현율 값 : 85 / (13 + 85)

4) 앙상블의 대표적인 방법 취합과 부스팅에 대해서 설명하고 각 방법에 포함되는 사이킷런의 클래스를 2개 이상 작성하시오.

- 취함 : 앙상블을 구현하는 각각의 모델들이 독립적으로 동작하며, 각 모델들의 상관관계가 0인 앙상블 구현 방법으로 각 하위 모델에 대해서 최대한 강하게 학습함
  ex) Voting, Bagging, RandomForest
- 부스팅 : 앙상블을 구현하는 각각의 하위 모델들이 선형으로 결합되어 점차적으로 성능을 향상시켜 나가는 방법으로 각 하위 모델에 대해서 강한 제약을 줌.
  ex) AdaBoosting, GradientBoosting

5) 이상 거래 인식을 위한 선형 분류 모델을 학습시킨 결과 거래금액, 거래지역, 시간대의 특성에 대해서 -7.515, 5.139, 0.345의 가중치 결과가 반환되었다. 해당 선형 모델의 예측 값 계산 공식을 작성하고 가중치 분석 결과를 작성하시오. (양성 : 이상 거래, 음성 : 정상 거래)

- y = [거래금액] * -7.515 + [거래지역] * 5.139 + [시간대] * 0.345
- 이상거래 인식을 위해서 사용한 컬럼 중 중요도의 순으로 나열하면 거래금액 > 거래지역 > 시간대 순으로 이해할 수 있음 이상 거래는 높은 거래 금액에서 발생될 가능성이 높고 또는 특정 거래 지역에서 발생할 가능이 높음

6) 머신러닝 모델의 학습 시 범주형(Categorical) 특성은 전처리 과정을 통해 모델의 학습 성능 및 일반화 성능을 향상시킬 수 있다. 범주형 특성의 전처리 방법 2가지를 설명하시오.

- 라벨 인코딩 : 범주형을 구성하는 데이터의 중복을 제거한 후, 각 데이터에 대해서 정수 값을 할당하고 전체 데이터를 정수 값으로 치환하는 전처리 방법
- 원핫 인코딩 : 범주형을 구성하는 데이터의 중복을 제거한 후, 각 데이터에 대해서 새로운 컬럼을 할당하여 다른 컬럼의 값은 0으로 해당 컬럼의 값은 1로 대치하는 전처리 방법

7) 머신러닝 모델의 학습에 사용되는 데이터를 학습과 테스트로 분할하는 목적과 기대효과에 대해서 설명하시오.

- 분할 목적 : 머신러닝 모델의 학습 이후 일반화 성능을 측정하기 위한 용도로 데이터를 학습과 테스트 데이터로 분할함
- 기대 효과 : 학습에서 사용되지 않은 테스트 데이터에 대한 성능을 바탕으로 미래의 데이터에 대한 기대치를 설정할 수 있음

8) 당신은 경비업체의 머신러닝 엔지니어이다. CCTV의 영상 데이터를 사용하여 위험 인물을 감지하는 역할을 수행하고 있다. 경비 비용 절감을 위해서는 머신러닝 모델이 인식한 인물이 위험인물일 확률이 높아야 하는 상황이다. 이러한 경우 분류 모델의 성능 지표 중 어느 것을 중요시해야 하는가?

- 위험 인물에 대한 정밀도 값

9) 분류 분석을 위한 머신러닝 모델을 개발하는 경우 데이터에서 클래스의 불균형 현상이 빈번하다. 클래스 불균형을 해결하기 위한 방법을 3개 이상 설명하시오.

1. 오버 샘플링 : 개수가 적은 클래스의 데이터를 개수가 많은 클래스의 데이터 개수와 유사한 수준으로 확대하는 방법 (단순복사, 통계 기반, 클러스터링 기반의 기법을 활용)
2. 언더 샘플링 : 개수가 많은 클래스의 데이터를개수가 작은 클래스의 데이터 개수와 유사하도록 랜덤 삭제
3. 가중치 조절 : 데이터 개수가 많은 클래스의 데이터 가중치를 감소시키고 데이터 개수가 적은 클래스의 가중치를 증가시켜 유사한 오차 값을 발생시키도록 제어하는 방법

10) 다음은 타이타닉 데이터에 대한 info 메소드 실행 결과이다. 결측 데이터가 존재하는 각 특성(컬럼)에 대해서 처리할 수 있는 방안과 그 이유를 설명하시오. (결측 데이터가 존재하는 각각의 특성에 대해서 처리 방안과 이유를 설명하시오)

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Id            1460 non-null   int64
 1   MSSubClass    1460 non-null   int64
 2   MSZoning      1460 non-null   object
 3   LotFrontage   1201 non-null   float64
 4   LotArea       1460 non-null   int64
 5   Street        1460 non-null   object
 6   Alley         91 non-null     object
 7   LotShape      1460 non-null   object
 8   LandContour   1460 non-null   object
 9   Utilities     1460 non-null   object
 10  LotConfig     1460 non-null   object
 11  LandSlope     1460 non-null   object
 12  Neighborhood  1460 non-null   object
dtypes: float64(1), int64(3), object(9)
memory usage: 148.4+ KB
```

- 결측치 컬럼명 : LotFrontPage
- 결측치 대처방안 : 행을 삭제 (결측치의 데이터 개수가 크지 않음), 지도학습 기반의 머신러닝을 활용하여 결측치를 대체, 비지도학습을 사용하여 군집분석으로 결측치를 대체

- 결측치 컬럼명 : Alley
- 결측치 대처방안 : 컬럼을 삭제 (결측이 아닌 데이터의 개수가 작아 학습에 큰 영향을 줄 수 없음)

< 작업형 >

1) house_prices.csv 데이터를 사용하여 회귀분석 모델을 구현하시오.
타겟특성 : SalePrice
요구사항
- 문자열 데이터의 경우 라벨 인코딩을 사용
- 수치 데이터의 경우 RobustScaler를 사용
- 결측 데이터가 존재하는 경우 해당 특성(컬럼)은 제외함
- Random Seed 값은 본인의 학번 가장 앞과 가장 뒤를 나열한 2자리로 설정함
  (EX: 학번이 20220422 -> 22)
- 머신러닝 클래스 3개를 사용하여 가장 좋은 모델 1개를 선정하고 이유를 2줄 내로 서술
- 모델 구축 후 평가 함수 mean_absolute_error을 사용하여 모델링 결과를 3줄 내로 서술

2) spaceship_titanic.csv 데이터를 사용하여 분류분석 모델을 구현하시오.
타겟 특성 : Transported
요구사항
- 문자열 데이터의 경우 원핫인코딩을 사용
  (타겟 데이터인 Transported 컬럼에 대해서 라벨 인코딩 수행)
- 수치 데이터의 경우 MinMaxScaler를 사용
- 결측 데이터가 존재하는 경우 해당 레코드는 제외함
- Random Seed 값은 본인의 학번 가장 앞과 가장 뒤를 나열한 2자리로 설정함
  (EX: 학번이 20220422 -> 22)
- 머신러닝 클래스 3개를 사용하여 가장 좋은 모델 1개를 선정하고 이유를 2줄 내로 서술
- 모델 구축 후 평가 함수 precesion_score을 사용하여 모델링 결과를 3줄 내로 서술
  (positive value는 1로 계산)